Appendix: Multimodal Omics Data Integration Using Max Relevance-Max Significance Criterion

Pradipta Maji and Ankita Mandal

***Abstract*—This paper presents a novel supervised regularized canonical correlation analysis, termed as CuRSaR, to extract relevant and significant features from multimodal high dimensional omics data sets [1]. The proposed method extracts a new set of features from two multidimensional data sets by maximizing the relevance of extracted features with respect to sample categories and significance among them. It integrates judiciously the merits of regularized canonical correlation analysis (RCCA) and rough hypercuboid approach. An analytical formulation, based on spectral decomposition, is introduced to establish the relation between canonical correlation analysis (CCA) and RCCA. It makes the computational complexity of the proposed algorithm significantly lower than existing methods. The concept of hypercuboid equivalence partition matrix of rough hypercuboid is used to compute both relevance and significance of a feature. The equivalence partition matrix also offers an efficient way to find optimum regularization parameters employed in CCA. The superiority of the proposed algorithm over other existing methods, in terms of computational complexity and classification accuracy, is established extensively on real life data.**

***Index Terms*—Multimodal data analysis, canonical correlation analysis, feature extraction, rough sets, classification.**

- **Computation of B.632+ Accuracy**

The $B.632+$ accuracy is defined as

$$B.632+ \text{ accuracy} = (1 - B.632+ \text{ error}); \tag{1}$$

where $B.632+$ error rate is defined as follows [2]:

$$B.632+ \text{ error} = (1 - \tilde{\omega})AE + \tilde{\omega}B1 \tag{2}$$

where AE denotes the proportion of the original training samples misclassified, termed as apparent error rate, and $B1$ is the bootstrap error, defined as follows:

$$B1 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\sum_{k=1}^{\mathrm{M}} I_{jk}Q_{jk}}{\sum_{k=1}^{\mathrm{M}} I_{jk}}\right) \tag{3}$$

where n is the number of original samples and M is the number of bootstrap samples. If the sample x_j is not contained in the kth bootstrap sample, then $I_{jk} = 1$, otherwise 0. Similarly, if x_j is misclassified, $Q_{jk} = 1$, otherwise 0. The weight parameter $\tilde{\omega}$ is given by

$$\tilde{\omega} = \frac{0.632}{1 - 0.368r};\ r = \frac{B1 - AE}{\gamma - AE};\ \gamma = \sum_{i=1}^{c} p_i(1 - q_i); \tag{4}$$

where c is the number of classes, p_i is the proportion of the samples from the ith class, and q_i is the proportion of them assigned to the ith class. Also, γ is termed as the no-information error rate that would apply if the distribution of the class-membership label of the sample x_j did not depend on its feature vector. In the current study, the SVM is used to compute different types of error rates. In order to calculate the $B.632+$ accuracy or error rate, apparent error (AE) is first calculated. This error is obtained when the same original data set is used to train and test a classifier. After that, the $B1$ error is computed from 100 bootstrap samples. Finally, the no-information error (γ) is calculated by randomly perturbing the class label of a given data set. The permutation of the class label is done 50 times. Each mutated data is used for feature extraction and the selected feature set is used to build the SVM. Then, the trained SVM is used to classify the original data set. The error generated by this procedure is known as γ rate. Finally, the $B.632+$ error and accuracy are computed based on AE, $B1$ error, and γ error using (2) and (1).

- **Computation of Accuracy and F1 Score**

The classification accuracy is defined as [3]

$$\text{accuracy} = \frac{\text{TP + TN}}{\text{TP + FP + TN + FN}}; \tag{5}$$

where TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative samples, respectively. Similarly, the F1 score is defined as follows:

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision + recall}}; \tag{6}$$

where precision and recall (sensitivity) are given by

$$\text{precision} = \frac{\text{TP}}{\text{TP + FP}}; \quad \text{and} \quad \text{recall} = \frac{\text{TP}}{\text{TP + FN}}. \tag{7}$$

This work is partially supported by the Department of Electronics and Information Technology, Government of India (PhD-MLA/4(90)/2015-16).

The authors are with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: {pmaji,amandal}@isical.ac.in.

Digital Object Identifier 10.1109/TBME.2016...

REFERENCES

[1] P. Maji and A. Mandal, "Multimodal Omics Data Integration Using Max Relevance-Max Significance Criterion," *IEEE Transactions on Biomedical Engineering*, vol. XX, no. YY, pp. 1–11, 2016.

[2] B. Efron and R. Tibshirani, "Improvements on Cross-Validation: The .632+ Bootstrap Method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 548–560, 1997.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, 1999.