Appendix: Discriminative Deep Canonical Correlation Analysis for Multi-View Data

Debamita Kumar and Pradipta Maji *

A1 Data Set Description

In this study, five cancer data sets, which include cervical carcinoma (CESC), colorectal carcinoma (CRC), kidney carcinoma (KIDNEY), lower grade glioma (LGG), and lung carcinoma (LUNG), and seven benchmark databases, namely, Digits [4], Caltech [5], CiteSeer [11], Cora [11], NW-OBJECT [2], Reuters [1], and ORL [14] are considered to study the performance of different algorithms. The detailed description of each of the data sets is presented here.

A1.1 Omics Data Sets

This section presents a brief description of the five multimodal omics data sets of The Cancer Genome Atlas (TCGA) [12] used in this work. All the data sets have been downloaded from the Genomic Data Commons Data Portal (http://cancergenome.nih.gov).

A1.1.1 Cervical Carcinoma (CESC)

This cancer accounts for 528,000 new cases and 266,000 deaths worldwide each year, more than any other gynecological tumour [3]. By comprehensive integrated analysis, TCGA research network has identified three subtypes in CESC [9]. The CESC data set consists of 104 samples: 32 samples of keratin-low squamous subgroup, 46 samples of keratin-high squamous subgroup, and 26 samples of adenocarcinoma-rich subgroup.

A1.1.2 Colorectal Carcinoma (CRC)

It is the third most commonly diagnosed cancer in both men and women and account for nine percent of all cancer deaths [6]. The colon and rectum are parts of the digestive system and cancer forms in the colon and/or the rectum. There are 261 samples in the CRC data set. Depending on the site of origin, the samples of CRC are divided into two subtypes, namely, colon carcinoma and rectum carcinoma, having 192 and 69 samples, respectively.

A1.1.3 Kidney Carcinoma (KIDNEY)

The kidney cancer data set has two histological subtypes, namely, renal clear cell carcinoma (KIRC) and renal papillary cell carcinoma (KIRP). These subtypes were included in the 2004 World Health Organization (WHO) classification of adult renal tumors [10]. The KIDNEY data consists of 305 samples of kidney cancer with 95 samples of KIRC and 210 samples of KIRP.

A1.1.4 Lower Grade Glioma (LGG)

Diffuse low-grade and intermediate-grade gliomas which together make up the lower grade gliomas have highly variable clinical behaviour that is not adequately predicted based on histological class. Integrative analysis of data from RNA, DNA copy number, and DNA methylation platforms has uncovered three prognostically significant subtypes of lower-grade glioma [8]. The LGG data set consists of 374 samples. The first subtype has 180 samples that exhibit IDH mutation and no 1p/19q codeletion. The second subtype exhibits both IDH mutation and 1p/19q codeletion and has 129 samples. The third one is called the wild-type IDH subtype and has 65 samples.

A1.1.5 Lung Carcinoma (LUNG)

Based on the same primary site of origin, the lung cancer set can be categorized into two subtypes, namely, adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). These were also the two major subtypes of lung cancer in 2015 WHO classification [13]. The LUNG data set consists of 546 samples with 312 samples of LUAD and 234 samples of LUSC.

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

These subtypes are clinically relevant and provide a roadmap for patient stratification and trials of targeted therapies. While CESC and CRC databases have four modalities, namely, DNA methylation (mDNA), protein expression (Protein), microRNA sequence (microRNA), and gene expression (RNA), KIDNEY, LGG, and LUNG data sets have five modalities, which are mDNA, Protein, microRNA, RNA, and copy number segmentation (CNS).

Data Platforms and Preprocessing: For the DNA methylation modality, methylation β-values from Illumina Human Methylation 450 platform is used. The Human Methylation 450 gives methylation β- values of approximately 450,000 CpG sites. Additionally, CpG locations with missing gene information were filtered out from the study. The top 2,000 most variable CpG sites are used in the current research problem. For the protein modality of all the data sets, reverse phase protein array data from the MDA_RPPA_Core platform is used. The number of proteins is different for each sample. Only a set of common proteins which is present in all the samples is considered to construct the protein expression data set. The miRNA sequence data is log transformed. The expression values of this modality are not available for most of the samples in the data sets. To avoid considering features with too many missing values, for all the omic modalities, those features for which the corresponding omic expression value is not present for more than 5% of the total number of samples are excluded. For the remaining features, missing values are replaced using 0. For the RNA modality of all the data sets, RNA-sequence data from the Illumina HiSeq platform is used which contains normalized RPKM (reads per kilobase of exon per million) counts for 20,531 genes. The data is then log transformed and 2,000 most variable genes based on their expression profile across the samples are considered. For KIDNEY, LGG, and LUNG data sets, CNS data from Affymetrix SNP Array 6.0 platform is used. The raw copy number segmented data is processed using the CNregions function of iCluster+ [7] R-package to reduce the redundant copy number regions. The CNregions function has an epsilon parameter which denotes the maximum Euclidean distance between adjacent probes tolerated for defining a non-redundant region. The number of non-redundant copy number regions extracted for a data set depends on the value of the epsilon parameter and is proportional to the number of samples in the data set. It is recommended in [7] to choose a value of epsilon such that the reduced dimension is less than 10,000. The default value of 0.005 is considered for the epsilon parameter of the CNregions function for the data sets.

A1.2 Benchmark Databases

A detailed description of the benchmark databases is presented in this section.

A1.2.1 Digits

The data has been downloaded from https://archive.ics.uci.edu/ml/datasets/Multiple+Features. The set consists of features of handwritten numerals ('0'-'9') extracted from a collection of Dutch utility maps. For each class, 200 patterns have been digitized in binary images which make a total of 2000 images in the database. These images of the digits are represented in terms of the following six feature sets:

1. mfeat-pix: 240 pixel averages in 2 x 3 windows.
2. mfeat-fou: 76 Fourier coefficients of the character shapes.
3. mfeat-fac: 216 profile correlations.
4. mfeat-zer: 47 Zernike moments.
5. mfeat-kar: 64 Karhunen-Love coefficients.
6. mfeat-mor: 6 morphological features.

It is to be mentioned here that for this database, the view mfeat-mor, consisting of 6 features, is not taken into consideration for the CCA based methods, since it does not met the minimum feature criteria with respect to the output number of features for the computation of canonical variables. Similar arguments hold for discriminant analysis based methods. However, deep learning based methods do not impose such conditions on input dimension of the modalities. Thus, in case of deep learning based methods, all six modalities are considered to extract features from the set.

A1.2.2 Caltech

This data has been downloaded from http://www.vision.caltech.edu/Image_Datasets/Caltech101. Caltech-101 consists of pictures of objects belonging to 101 categories. There are 40 to 800 images per category. Most categories have about 50 images. Collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc Aurelio Ranzato. The size of each image is roughly 300 x 200 pixels. The 2386 images of the database are represented in terms of the following six feature sets:

1. 48 Gabor features
2. 40 Wavelet moments

3. 254 CENTRIST features
4. 1984 HOG features
5. 512 GIST features
6. 928 local binary patterns

Caltech-20 is a subset of Caltech-101, which contains only 20 classes. In the current research work, Caltech-20 is used to analyze the performance of the proposed model as well as existing approaches.

A1.2.3 CiteSeer

The CiteSeer database is obtained from http://networkrepository.com. The set is generated by sampling scientific documents from CiteSeer digital library. The publications are classified into one of the six classes, namely, Agents, AI, DB, IR, ML, and HCI. There are 3312 papers in the data set. The papers are selected in a way such that in the final set every paper cites or is cited by atleast one other paper. After stemming and removing the stopwords, a vocabulary of size 3703 unique words is obtained. All the words with document frequency less than 10 are removed. Each publication in the database is described by a 0 or 1 valued word vector indicating the absence or presence of the corresponding word in the document.

A1.2.4 Cora

This is a standard benchmark dataset of research articles, downloaded from http://networkrepository.com. The Cora data set consists of machine learning papers. These papers are classified into one of the seven topics, which include Case_Based, Genetic_Algorithms, Neural_Networks, Probabilistic_Methods, Reinforcement_Learning, Rule_Learning, and Theory. The articles are selected in a way such that in the final set every paper cites or is cited by atleast one other paper. There are 2708 papers in the data set. After stemming and removing the stopwords, a vocabulary of size 1433 unique words is obtained. All the words with document frequency less than 10 are removed. Each article in the database is described by a 0 or 1 valued word vector indicating the absence or presence of the corresponding word in the document.

A1.2.5 NW-OBJECT

The NW-OBJECT data set, which is formally referred to as NUS-WIDE-OBJECT, is a subset of NUS-WIDE database and has been downloaded from https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html. This database, created by Lab for Media Search in National University of Singapore, is intended for object recognition task. It consists of 30000 images, categorized into 31 different classes. The 30000 images of the database are represented in terms of the following five feature sets:

1. 64 color histogram features
2. 144 color correlogram features
3. 75 edge direction histogram features
4. 128 texture wavelet features
5. 225 block-wise color moment features

A1.2.6 Reuters

This multilingual data has been obtain from http://archive.ics.uci.edu/ml/machine-learning-databases/00259. The collection contains feature characteristics of documents originally written in English language and the corresponding translations in French, German, Spanish, and Italian languages over a common set of 6 categories. This collection can be used for multilingual categorization and multiview learning research. Documents have been translated, preprocessed, and are made available as feature characteristics in a "bag of words" format. 18758 documents are partitioned into 6 categories which include CCAT, C15, ECAT, E21, GCAT and M11, and represented in terms of the following five feature sets:

1. EN-EN : Original English documents with vocabulary size 21531
2. FR-EN : French documents translated to English with vocabulary size 24893
3. GR-EN : German documents translated to English with vocabulary size 34279
4. IT-EN : Italian documents translated to English with vocabulary size 15506
5. SP-EN : Spanish documents translated to English with vocabulary size 11547

A1.2.7 ORL

This is a standard face recognition database, obtained from http://cam-orl.co.uk. The set contains images from 40 individuals, each providing 10 different images. For some subjects, the images were taken at different times. The facial expressions (open or closed eyes, smiling or non-smiling) and facial details (glasses or no glasses) also vary. The images were taken with a tolerance for some tilting and rotation of the face of up to 20 degrees. Moreover, there is also some variation in the scale of up to about 10 percent. All images are gray-scale and normalized to a resolution of 92 $\times$ 112 pixels, which are rescaled 64 $\times$ 64 pixels in the current study.

References

[1] M. R. Amini, N. Usunier, and C. Goutte. Learning from Multiple Partially Observed Views - an Application to Multilingual Text Categorization. *Advances in Neural Information Processing Systems*, 22:28–36, 2009.

[2] T-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y-T. Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore, 2009.

[3] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns. *International Journal of Cancer*, 136(5):E359–E386, 2015.

[4] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[5] F. Li, R. Fergus, and P. Perona. One-Shot Learning of Object Categories. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[6] I. Mármol, C. Sánchez de Diego, A. P. Dieste, E. Cerrada, and M. J. R. Yoldi. Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *International Journal of Molecular Sciences*, 18(1):197, 2017.

[7] Q. Mo and R. Shen. iClusterPlus: Integrative Clustering of Multiple Genomic Data Sets. 2013.

[8] Cancer Genome Atlas Research Network. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.

[9] Cancer Genome Atlas Research Network. Integrated Genomic and Molecular Characterization of Cervical Cancer. *Nature*, 543(7645):378, 2017.

[10] S. R. Prasad, P. A. Humphrey, J. R. Catena, V. R. Narra, J. R. Srigley, A. D. Cortez, N. C. Dalrymple, and K. N. Chintapalli. Common and Uncommon Histologic Subtypes of Renal Cell Carcinoma: Imaging Spectrum with Pathologic Correlation. *Radiographics*, 26(6):1795–1806, 2006.

[11] R. Rossi and N. Ahmed. The Network Data Repository With Interactive Graph Analytics and Visualization, 2015.

[12] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemporary Oncology*, 19(1A):A68, 2015.

[13] W. D. Travis, E. Brambilla, A. P. Burke, A. Marx, and A. G. Nicholson. Introduction to the 2015 World Health Organization Classification of Tumors of the Lung, Pleura, Thymus, and Heart. *Journal of Thoracic Oncology*, 10(9):1240–1242, 2015.

[14] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang. Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(1):131–137, 2004.