Similarity Based Canonical Correlation Analysis (SDCCA)

Pranay Garg (0801CS171056)

pranaygarg001@gmail.com

December 15, 2020

Contents

Chapter 1
Introduction

1.1 MultiView Learning

Multi-view learning (MVL) is a strategy for fusing data from different sources or subsets. Many important real-life problems such as video surveillance, medical diagnosis, fashion industry, sentiment analysis, etc. can not be handled well through a single view. This is because a single view contains only partial information related to the problem. Thus consideration of data from multiple views of the same problem is essential.Few researchers have classified the existing MVL approaches into three major categories: co-training, multi-kernel learning and subspace learning representation. One of the classical approaches in MVL is canonical correlation analysis (CCA) first proposed by Hotelling.It is a subspace learning representation approach.

1.2 SDCCA

CCA finds pairs of projections from different views, such that the correlation between the transformed views in the common subspace is maximized. CCA looks for latent low dimensional information from two views of common subspace. In real-world applications, data from multiple views often contain similar or complementary information. In general, most of the proposed CCA based approaches do not exploit the similarity and complementary information. Therefore existing approaches are not able to build comprehensive latent space for multiview learning. As a result, existing multiview learning algorithms do not perform well and fail to produce correct results. To address this issue, in this paper, a canonical correlation analysis with Bhattacharya similarity distance for multiview data representation (SDCCA) is proposed.

Chapter 2
Mathematical Formulation

2.1 Formulation

Through this method, latent space for multiview learning is determined by incorporating similarities and complementary information from multiple views with cross correlation analysis. It also explores the local manifold structures within the view. SDCCA objective function is directly developed from the original objective function of CCA.

Let a set of pair-wise sample data (X,Y) for multiview learning where X and Y are represented by $X = [x_1 , x_2 , x_3 \ldots\ldots x_n] \epsilon \mathbb{R}^{d_x}$ and $Y = [y_1 , y_2 , y_3 \ldots\ldots y_n] \epsilon \mathbb{R}^{d_y}$ with n instances and d_x and d_y dimension respectively. To target the goal, the objective function of SDCCA is given in the following form:

$$\max_{w_x, w_y} \frac{W_x^T \tilde{c}_{xy} W_y}{\sqrt{W_x^T c_{xx} W_x} \sqrt{W_y^T c_{yy} W_y}}$$

$$s.t. \quad W_x^T c_{xx} W_x = 1, W_y^T c_{yy} W_y = 1$$

Where

$$\tilde{c}_{xy} = \sum_{i=1}^{n} x_i y_i^T + \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^x x_i y_j^T + \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^y x_i y_j^T + \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^{xy} x_i y_j^T$$

$$= XY^T + Xs^x Y^T + Xs^y Y^T + Xs^{xy} Y^T$$

$$= X(I + s^x + s^y + s^{xy})Y^T$$

Here c_{xx} and c_{yy} are the XX^T and YY^T respectively. s^x , s^y are the neighbourhood information between samples defined by,

$$s_{ij}^{x} = \begin{cases} exp(-\|x_i - x_j\|^2/t_x) & if\ x_j \epsilon NE(x_i)\ or\ x_i \epsilon NE(x_j) \\ 0 & \text{otherwise} \end{cases}$$

$$s_{ij}^{y} = \begin{cases} exp(-\|y_i - y_j\|^2/t_y) & if\ y_j \epsilon NE(y_i)\ or\ y_i \epsilon NE(y_j) \\ 0 & \text{otherwise} \end{cases}$$

s^{xy} is a dissimilarity between samples and it is defined as $1 - s^{\sim xy}$, where $s^{\sim xy}$ is similarity between views x and y. It is calculated using Bhattacharya similarity coefficient defined as:

$$\tilde{s}^{xy} = cos(\theta) = \sum_{i \epsilon N} \sqrt{x_i . y_i}$$

The optimization problem of SDCCA can be rewritten as the Lagrangian optimization as follows:

$$\begin{bmatrix} & \tilde{c}_{xy} \\ \tilde{c}_{xy} & \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} \tilde{c}_{xy} & \\ & \tilde{c}_{xy} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}$$

This can be efficiently computed via the well known SVD to obtain the projective vector w_x and w_y.

2.2 Universe Subspace

CCA is viewed as learning a common subspace such that correlation between two data views is maximized. The common subspaces between two views can be formulated as

$$C = \frac{Xw_x + Yw_y}{2} \qquad \text{Or} \qquad U = \frac{1}{m} \sum_{v=1}^{m} (X_v w_v)$$

Where m is total views and U is universe subspace of all views.

Chapter 3
Algorithm

3.1 SDCCA Algorithm

Algorithms 1: SDCCA algorithm

Input : multiview training datasets $X \epsilon \mathbb{R}^{nxd_1}$,$Y \epsilon \mathbb{R}^{nxd_2}$
where d_1 and d_2 are the dimensionality of X and Y views.
Output: w_x and w_y

1. Construct $S^x, S^y and S^{xy}$:
   - 1.1 if x_i and y_j are in neighbors within the t_x kernel windows
     $S^x = exp(-\|x_i - x_j\|^2/t_x)$; otherwise 0.
   - 1.2 if y_i and x_j are in neighbors within the t_y kernel windows
     $S^y = exp(-\|y_i - y_j\|^2/t_y)$; otherwise 0.
   - 1.3 $s^{xy} = 1$ - $\sum_{i \epsilon N} \sqrt{x_i.y_i}$.
2. Define S = $I + S^x + S^y + S^{xy}$
3. Co-variance matrix:
   $\tilde{C}_{xy} = XSY^T$
5. Compute C_{xx} *and* C_{yy}
6. Compute matrix H =$C_{xx}^{-\frac{1}{2}} \tilde{C}_{xy} C_{yy}^{-\frac{1}{2}}$
7. Perform SVD decomposition on H: H=UDV^T.
8. Obtain $w_x = C_{xx}^{-\frac{1}{2}} U$ *and* $w_y = C_{yy}^{-\frac{1}{2}} V$

Chapter 4
Documentation of API

4.1 Package organization

from SDCCA import SDCCA
class SDCCA(scale = True,k=2)

Parameters

- scale (boolean) (default true) : whether to scale the data or not
- algorithm (string : “k-means” or “2-norm”) (default k-means) (just an idea for future releases)
- k (integer) (default 2) : number of neighbours to consider in K-Means algorithm.

Attributes

- x : first data set of dimension n*p with n samples and p features.
- y : second data set of dimension n*q with n samples and q features.
- wx , wy : final projection vectors of two views of p*p and q*q.
- n : number of samples in datasets
- p : number of features in x dataset
- q : number of features in y dataset

Methods

- fit(x,y)
  - Learn and apply SDCCA on the data and find projection vectors.
  - Parameters : x and y are dataframes of dimensions n*p and n*q respectively.
  - Returns the two projection vectors w_x and w_y of views x, y respectively.

- fit_transform(x,y)
  - First finds projection vectors and then returns the projected views after applying projection vectors.
  - Parameters : x and y are dataframes of dimensions n*p and n*q respectively.
  - Returns the two projected views x,y after applying projection vectors w_x and w_y respectively.

Chapter 5
Examples

5.1 Example 1

```
from SDCCA import SDCCA
import pandas as pd
x = np.array([[0., 0., 1.], [1.,0.,0.], [2.,2.,2.], [3.,5.,4.]])
y = np.array([[0.1, 0.2], [0.9, 1.1], [6.2, 5.9], [11.9, 12.3]])
sdcca = SDCCA(scale = False) # k is by default 2
wx, wy = sdcca.fit(x,y) # projection vectors
```

5.2 Example 2

```
from SDCCA import SDCCA
import pandas as pd
x = np.array([[1., 2.1, 1.], [1., 2.1, 0.5], [2.,2.,2.], [3.,5.,4.]])
y = np.array([[0.1, 4.2], [0.9, 2.1], [7.2, 5.9], [12.9, 1.3]])
sdcca = SDCCA(k=4) # scale is by default true
x_new, y_new = sdcca.fit_transform(x,y) # projected views
```

Chapter 6
Learning Outcome

6.1 Multiview Learning

- With the help of this project I learned how to handle multiple views together and find correlation between them using CCA.
- Also learned how multiple views of the same dataset can be used to solve more complex problems.
- Understood concepts and overcame barriers much like real world scenarios.

6.2 Others

- Strengthened the concepts of python programming language.
- Understood the art of reading research papers and understanding notations.
- Also got familiar with open source contribution.

Appendix A

References

1. Surendra Gupta, Urjita Thakar, and Sanjiv Tokekar : Canonical Correlation Analysis with Bhattacharya Similarity Distance for Multiview Data Representation Research Paper .
2. Chenfeng Guo and Dongrui Wu : Canonical Correlation Analysis (CCA) Based Multi-View Learning: An Overview Research Paper .
3. Zhang D. Wang F.: A new locality-preserving canonical correlation analysis algorithm for multi-view dimensionality reduction, Neural Process Lett., vol. 37, pp. 135-146, July 2012. doi:10.1007/s11063-012-9238-9
4. A. Bhattacharyya.: On a measure of divergence between two multinomial populations, SankhyA: The Indian Journal of Statistics, vol. 7(4), pp. 401-406, 1946. ¨ Research Paper .