Semi Supervised Multi-View Correlation Feature Learning with Application to Webpage Classification

Team - 3 (Abnormal Distribution)
Shobhan Mandal(20172064)
Arijit Mukherjee(20172081)
Ranajit Saha(20172119)
Sanjoy Chowdhury(20172123)

Semi Supervised Learning:

- ***Simple Terms:*** A class of supervised learning tasks and techniques that make use of unlabelled data for training - typically a small amount of labelled data with large amount of unlabelled data. (It lies in between the un-supervised learning and supervised learning).

- Here, Labeled data is used to help identify *that* there are specific groups of webpage types present in the data-set and what they *might be.* The algorithm is then trained on unlabeled data to define the boundaries of those web page types and may even identify new types of web-pages that were unspecified in the existing human inputted labels.

Multi-view Learning :

- Many real-world datasets possess data samples characterized using multiple views,e.g., web-pages can be described using both textual content in each page and the hyperlink structure between them.
- It has been shown that the error rate on unseen test samples can be upper bounded by the disagreement between the classification decisions obtained from independent views of the data.

This relatively new machine learning technique, commonly called as **Multiview Learning** has been predominantly successfully used in conjunction with semi-supervised and also unsupervised approaches.

- Each example is described using two different feature sets that provide different, complementary information about the instance.
- Ideally, the two views are conditionally independent (i.e., the two feature sets of each instance are conditionally independent given the class) and each view is sufficient (i.e., the class of an instance can be accurately predicted from each view alone).
- Co-training first learns a separate classifier for each view using any labeled examples. The most confident predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data.

Introduction

Webpage classification refers to the problem of assigning a webpage class that describes its contents. A webpage classification has 3 characteristics :

- Webpage is a kind of multi-view data, since it usually contains two or more types of data, e.g., text, hyperlinks and images, where each type of data can be regarded as a view.

- Webpage classification is a semi-supervised application, since labeled pages are harder to collect compared to unlabeled pages in practice.

- Webpage data is high-dimensional, since webpages usually contain much information. Considering these three characteristics, it is crucial to design effective semi-supervised multi-view feature learning (SMFL) methods for webpage classification.

These 3 characteristics have been taken into account by two other webpage classification methods only :-

- SSGCA
- USI^{2}MD

What this paper is about...

- Webpage classification application is usually semi-supervised and currently using **Semi-supervised Multi-view Feature Learning** (SMFL) technique to deal with the webpage classification problem is on the rise.

- Effectively utilize the correlation information among multi-view of webpage data is an important research topic. Correlation analysis on multi-view data can facilitate extraction of the complementary information.

- This paper proposes a novel SMFL approach **Semi-supervised Multi-view Correlation Feature Learning** (**SMCFL)**, for web-page classification.

- SMCFL seeks for a discriminant common space by learning a multi-view shared transformation in a semi-supervised manner.

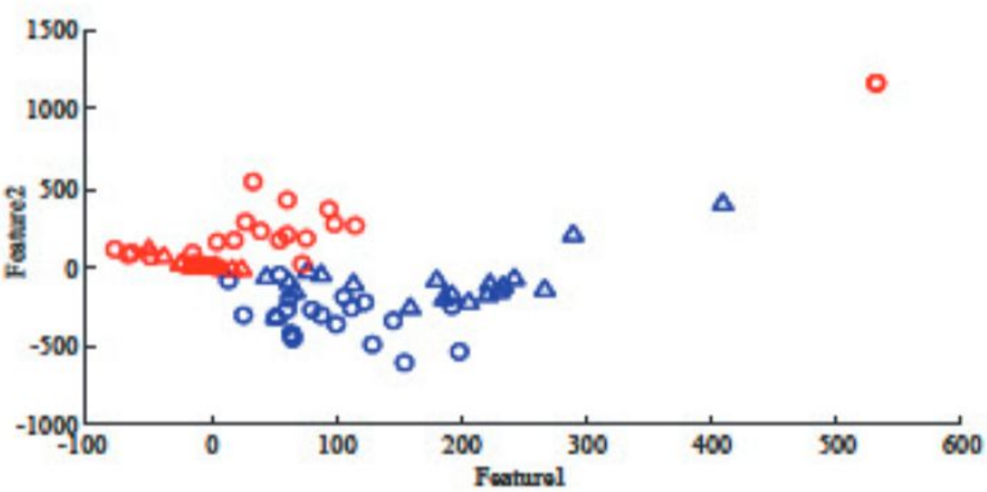
- In the discriminant space, the correlation between intra-class samples is maximized, and the correlation between inter-class samples and the global correlation among both labeled and unlabeled samples are minimized simultaneously.

Motivation & Contribution

- The correlation information from inter-view and intra-view depicts the association relationship among multiple views which has close connection with classification.
- The “inter-view” and “intra-view” mean the relationship between samples across different views and within certain view, respectively.
- The inter-view correlation contains the within-class and between-class correlation of samples across different views.
- The intra-view correlation contains the within-class and between-class samples within the same view.

From **WebKB** dataset as an example, we randomly select 10 webpage samples of each class in the link and page views and perform the Principal Component Analysis (PCA) transformation to obtain two major components of each sample for plotting the sample distribution in the graph next page.

Motivation & Contribution (Contd.)



Sample Distribution of 40 webpage samples in the **WebKB** dataset, where **RED** and **BLUE** colors *denote two views*, and
CIRCLE(○) and TRIANGLE (△) markers *denote sample points from two classes*.

Here we are finding that samples of different views own small correlation and between-class samples within each view own relatively large correlation.

Therefore we would be maximizing the correlation of samples from different views while in the same class and minimize the correlation of samples in the same view wule from different classes to make the distribution more favourable for classification.

Motivation & Contribution (Contd.)

The most noted drawbacks in the current SMFL methods are:-

- Some of the Existing SMFL methods **do not consider the correlation information**
- For other there exists much room to **improve their discriminant abilities,** since
    - They **do not explore intra-view correlation information**.
    - They **simply maximizes** the **correlation** from both **intra-view** and **inter-view**, which **cannot** make **full use** of the **discriminant correlation** from the aspect of classification.
    - They **only maximizes** the **inter-view within-class** correlation without considering the **inter-view between-class** correlation.

The summarization of the study:

- The objective function of SMCFL is designed to maximize the correlation between intra-class samples, and minimize the correlation between inter-class samples and the global correlation among both labeled and unlabeled samples. SMCFL can thus effectively explore the intra-view and inter-view discriminant correlation information.

- The solution is global optimal and can be derived analytically without iterative calculation. (The authors believe the proposed work is the first to present this kind of solution, which can be applied to other correlation-based feature learning problems.)

- SMCFL is verified on a widely used webpage dataset (WebKB). The experimental results show that it can significantly outperform state-of-the-art webpage classification method.

Objective Function of SMCFL

- Suppose that X_l = $\{X_1,X_2,...,X_c\}$ is the labeled training webpage sample set from C classes, where each $X_i(i=1,...,C)$ contains webpage samples of M views and $x^s_{ip} \in R^{dx1}$ denotes the p^{th} webpage sample from the s^{th} view of the i^{th} class.
- Here, d denotes the dimensionality of samples.
- Assume that l^s_i denotes the number of samples from the s^{th} view and the i^{th} class, and l_i = $\sum_{s=1}^{M} l_i^s$ denotes the number of samples in the i^{th} class. Let X^M be the unlabeled training sample set, $X=\{X^l,X^M\}$, and N denotes the total sample number in X. For simplicity of representation, we regard X^M as the $(C+1)^{th}$ class.

$$S_w = \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}WW^Tx_{iq}^t}{\sqrt{x_{ip}^{sT}WW^Tx_{ip}^s}\sqrt{x_{iq}^{tT}WW^Tx_{iq}^t}}\right], \qquad (1)$$

$$S_b = \frac{2}{C(C-1)}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}WW^Tx_{jq}^t}{\sqrt{x_{ip}^{sT}WW^Tx_{ip}^s}\sqrt{x_{jq}^{tT}WW^Tx_{jq}^t}}\right], (2)$$

$$S_t = \frac{1}{(C+1)C}\left(\begin{array}{l}\displaystyle\sum_{i=1}^{C}\sum_{j=i+1}^{C+1}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}WW^Tx_{jq}^t}{\sqrt{x_{ip}^{sT}WW^Tx_{ip}^s}\sqrt{x_{jq}^{tT}WW^Tx_{jq}^t}}\right]\\ \displaystyle+\frac{C}{2}\sum_{i=1}^{C+1}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}WW^Tx_{iq}^t}{\sqrt{x_{ip}^{sT}WW^Tx_{ip}^s}\sqrt{x_{iq}^{tT}WW^Tx_{iq}^t}}\right]\end{array}\right). \quad (3)$$

Then, the objective function of SMCFL can be defined as

$$\max_{W}\ f(W) = S_w - r_1S_b - r_2S_t, \qquad (4)$$

where $r_1 > 0$ and $r_2 > 0$ are weight coefficients.

We set $H = WW^T$ and $||x^s_{ip}|| = 1 \quad (\forall i, s, p)$, where $||.||$ denotes the l_2-norm of a vector. Obviously, H should be symmetric, i.e., $H = H^T$

We relax (4) into the following formulation:

$$\max_{H} f(H) = S'_w - r_1 S'_b - r_2 S'_t \, ,$$
$$s\,t. \quad H = H^T$$

where S'_w, S'_b, and S'_t are separately defined as follows:

$$S'_w = \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}WW^T x_{iq}^t}{\left\|x_{ip}^s\right\|\left\|x_{iq}^t\right\|\left\|WW^T\right\|_F}\right], \qquad (6)$$
$$= \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}Hx_{iq}^t}{\left\|H\right\|_F}\right]$$

$$S'_b = \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}WW^T x_{jq}^t}{\left\|x_{ip}^s\right\|\left\|x_{jq}^t\right\|\left\|WW^T\right\|_F}\right], (7)$$
$$= \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}Hx_{jq}^t}{\left\|H\right\|_F}\right]$$

$$S'_t = \frac{1}{2(C+1)C}\left(\begin{array}{l}\displaystyle\sum_{i=1}^{C+1}\sum_{j=i,j\neq i}^{C+1}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}Hx_{jq}^t}{\left\|H\right\|_F}\right]\\ \displaystyle +C\sum_{i=1}^{C+1}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}Hx_{iq}^t}{\left\|H\right\|_F}\right]\end{array}\right). \qquad (8)$$

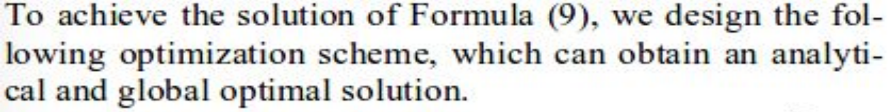
Similarly (5) can be further translated into:

$$\begin{aligned} &\min_{H} \|H\|_F \\ &s.t.\ \ B \geq 1,\ H = H^T,\ H \geq 0 \end{aligned} , \qquad (9)$$

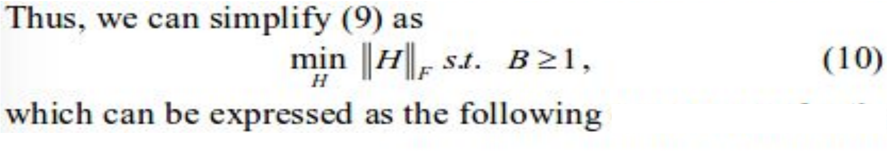
where

$$\begin{aligned} B = {} & \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t} x_{ip}^{sT} H x_{iq}^{t}\right] \\ & -\frac{r_1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t} x_{ip}^{sT} H x_{jq}^{t}\right] \\ & -\frac{r_2}{2(C+1)C}\left(\begin{aligned} & \sum_{i=1}^{C+1}\sum_{j=1,j\neq i}^{C+1}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t} x_{ip}^{sT} H x_{jq}^{t}\right] \\ & + C\sum_{i=1}^{C+1}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t} x_{ip}^{sT} H x_{iq}^{t}\right] \end{aligned}\right) \end{aligned} .$$

Solution of SMFCL



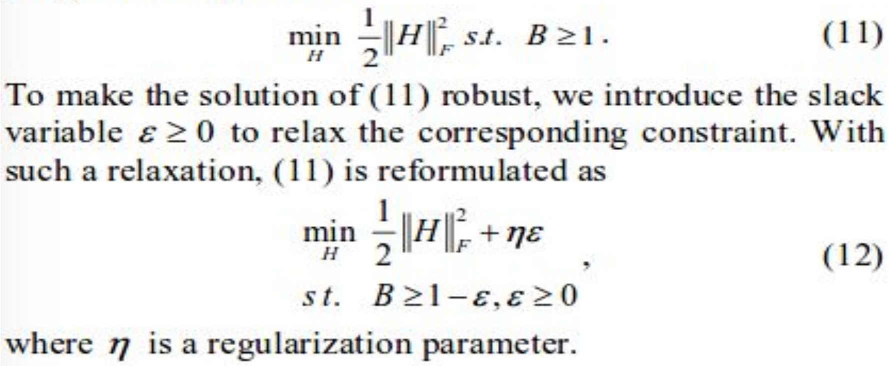
To achieve the solution of Formula (9), we design the following optimization scheme, which can obtain an analytical and global optimal solution.



Thus, we can simplify (9) as

$$\min_{H} \|H\|_F \; s.t. \quad B \geq 1, \tag{10}$$

which can be expressed as the following



$$\min_{H} \frac{1}{2}\|H\|_F^2 \; s.t. \quad B \geq 1. \tag{11}$$

To make the solution of (11) robust, we introduce the slack variable $\varepsilon \geq 0$ to relax the corresponding constraint. With such a relaxation, (11) is reformulated as

$$\begin{aligned} &\min_{H} \frac{1}{2}\|H\|_F^2 + \eta\varepsilon \\ &s.t. \quad B \geq 1-\varepsilon, \varepsilon \geq 0 \end{aligned}, \tag{12}$$

where η is a regularization parameter.

By applying the Lagrangian technique to constrained optimization problem, we define the Lagrange function as

$$L(H,\varepsilon,\alpha,\mu)=\frac{1}{2}\|H\|_F^2+\eta\varepsilon-\alpha(B-1+\varepsilon)-\mu\varepsilon\ , \qquad (13)$$

where α and μ are Lagrangian multipliers. By making the derivatives with respect to H and ε equal to zeros, we can obtain

$$\frac{\partial L}{\partial H}=H-\alpha R=0 \text{ and } \frac{\partial L}{\partial \varepsilon}=\eta-\alpha-\mu=0\,, \qquad (14)$$

where

$$\begin{aligned} R=&\frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}x_{ip}^{s}x_{iq}^{tT}\right]\\ &-\frac{r_1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}x_{ip}^{s}x_{jq}^{tT}\right]\\ &-\frac{r_2}{2(C+1)C}\left(\begin{aligned}&\sum_{i=1}^{C+1}\sum_{j=1,j\neq i}^{C+1}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}x_{ip}^{s}x_{jq}^{tT}\right]\\ &+C\sum_{i=1}^{C+1}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}x_{ip}^{s}x_{iq}^{tT}\right]\end{aligned}\right)\end{aligned} \quad . \qquad (15)$$

The corresponding Karush-Kuhn-Tucher (KKT) conditions (Chen et al. 2011) are

$$\alpha\left(B-1+\varepsilon\right)=0\ ,\ \mu\geq 0 \text{ and } \alpha\geq 0\ . \tag{16}$$

According to (14), we obtain

$$H=\alpha R\ . \tag{17}$$

It can be easily proved that R is a symmetric matrix, and thus $H=\alpha R$ is symmetric. Then, we can get

$$B=tr\left(HR^{T}\right), \tag{18}$$

where $tr\left(\cdot\right)$ denotes the trace of a square matrix.

Substituting Eqs. (14), (17) and (18) into (13), we obtain its Wolfe dual objective as follows

$$DL(H,\varepsilon,\alpha,\mu)=-\frac{\alpha^{2}}{2}tr\left(RR^{T}\right)+\alpha\ . \tag{19}$$

Hence, to get the solution of (12) is equivalent to solving the following optimization problem:

$$\max_{\alpha}\ \alpha-\frac{A}{2}\alpha^{2}, \tag{20}$$

where $A=tr\left(RR^{T}\right)$ is positive. (20) is a convex quadratic programming problem. If $\eta\geq 1/A$, the solution of (20) is $\alpha^{*}=1/A$; otherwise, the solution is $\alpha^{*}=\eta$.

Finally, substituting $\boldsymbol{\alpha}$ into (17), we can get H. To obtain the projective transformation matrix W, H is eigen-decomposed as

$$H = U \Lambda U^T, \tag{21}$$

where Λ is a diagonal eigenvalue matrix of H, and U is an orthogonal matrix whose columns correspond to the eigenvectors of H.

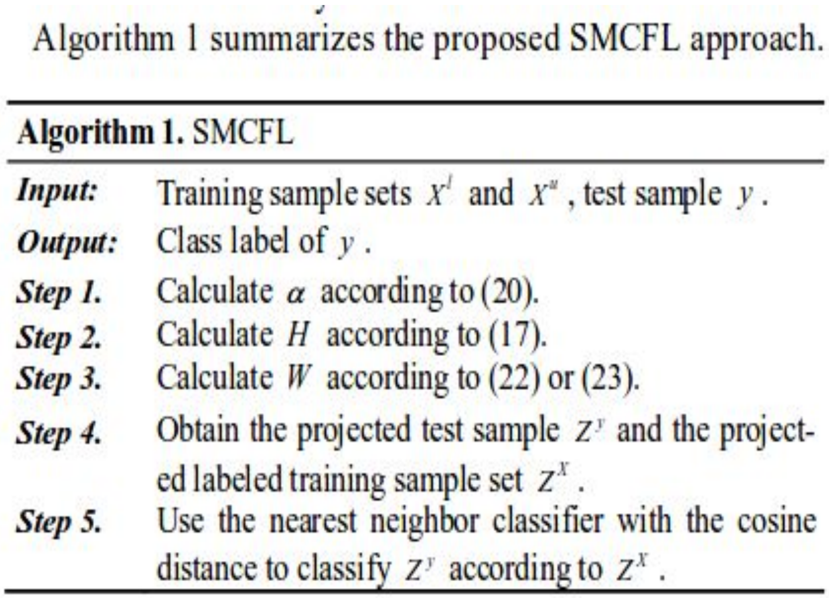
If H is positive semi-definite, we can obtain W by

$$W = U \sqrt{\Lambda}. \tag{22}$$

When H is not positive semi-definite, namely some eigenvalues of H are negative, like the solution trick in (Ma et al. 2007), we select the positive eigenvalues and corresponding eigenvectors to construct a new diagonal matrix Λ_+ ($\Lambda_+ = \sqrt{\Lambda_+}\left(\sqrt{\Lambda_+}\right)^T$) and a new orthogonal matrix U_+, respectively. Then we can obtain W by

$$W = U_+ \sqrt{\Lambda_+}. \tag{23}$$

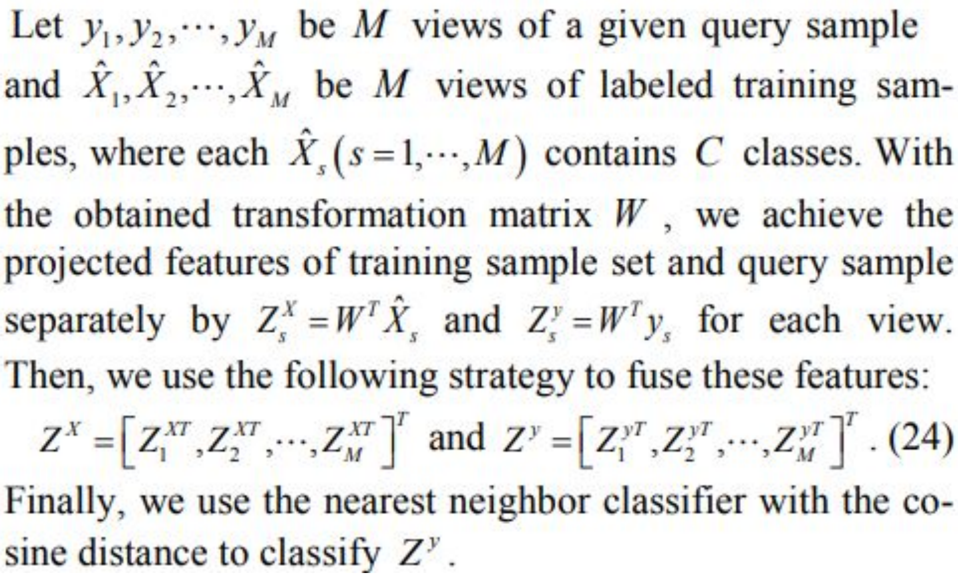
Solution of SMFCL (contd.) [Algorithm]



Algorithm 1 summarizes the proposed SMCFL approach.

Algorithm 1. SMCFL

```
Input:   Training sample sets $X^l$ and $X^u$, test sample $y$.
Output:  Class label of $y$.
Step 1.  Calculate $\alpha$ according to (20).
Step 2.  Calculate $H$ according to (17).
Step 3.  Calculate $W$ according to (22) or (23).
Step 4.  Obtain the projected test sample $Z^y$ and the projected labeled training sample set $Z^X$.
Step 5.  Use the nearest neighbor classifier with the cosine distance to classify $Z^y$ according to $Z^X$.
```

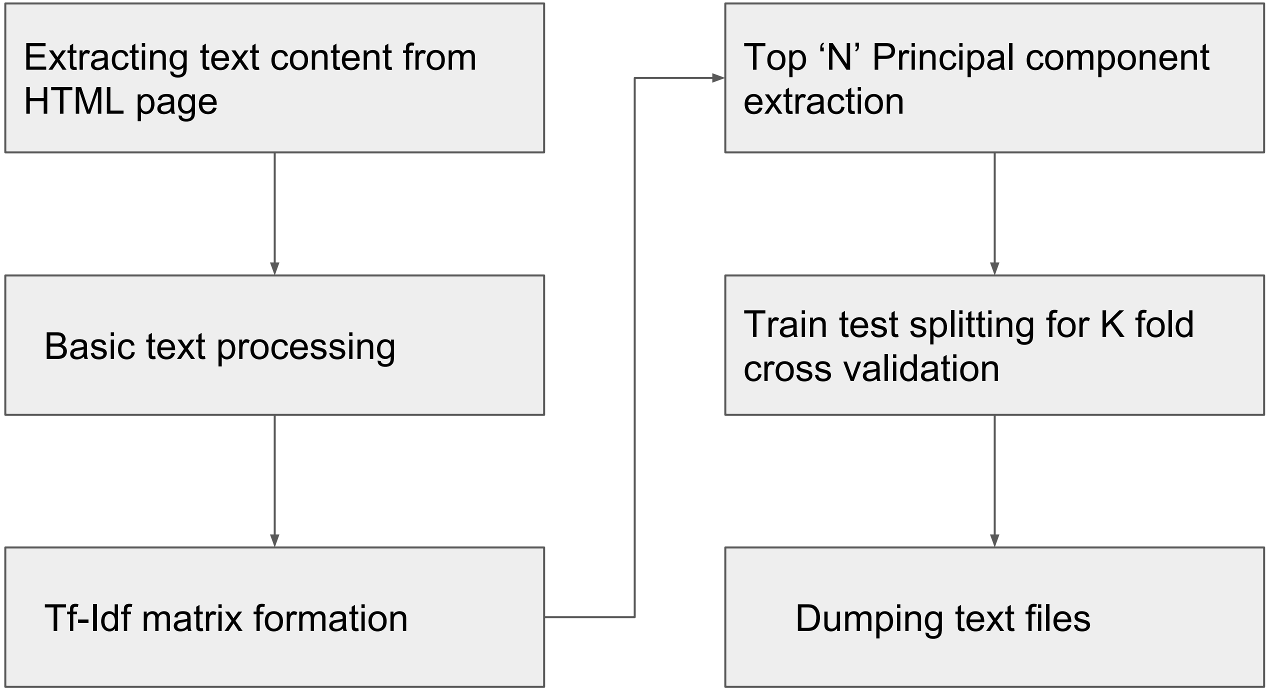


Let $y_1, y_2, \cdots, y_M$ be M views of a given query sample and $\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_M$ be M views of labeled training samples, where each $\hat{X}_s\,(s = 1, \cdots, M)$ contains C classes. With the obtained transformation matrix W, we achieve the projected features of training sample set and query sample separately by $Z_s^X = W^T \hat{X}_s$ and $Z_s^y = W^T y_s$ for each view. Then, we use the following strategy to fuse these features:

$$Z^X = \left[Z_1^{XT}, Z_2^{XT}, \cdots, Z_M^{XT}\right]^T \text{ and } Z^y = \left[Z_1^{yT}, Z_2^{yT}, \cdots, Z_M^{yT}\right]^T . \quad (24)$$

Finally, we use the nearest neighbor classifier with the cosine distance to classify Z^y.

Data Preprocessing



Extracting text content from HTML page

Basic text processing

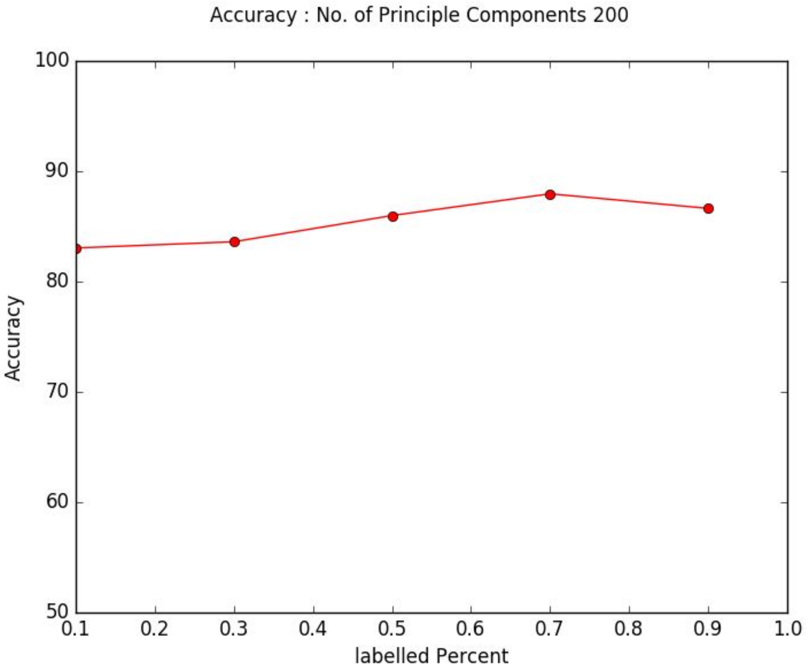
Tf-Idf matrix formation

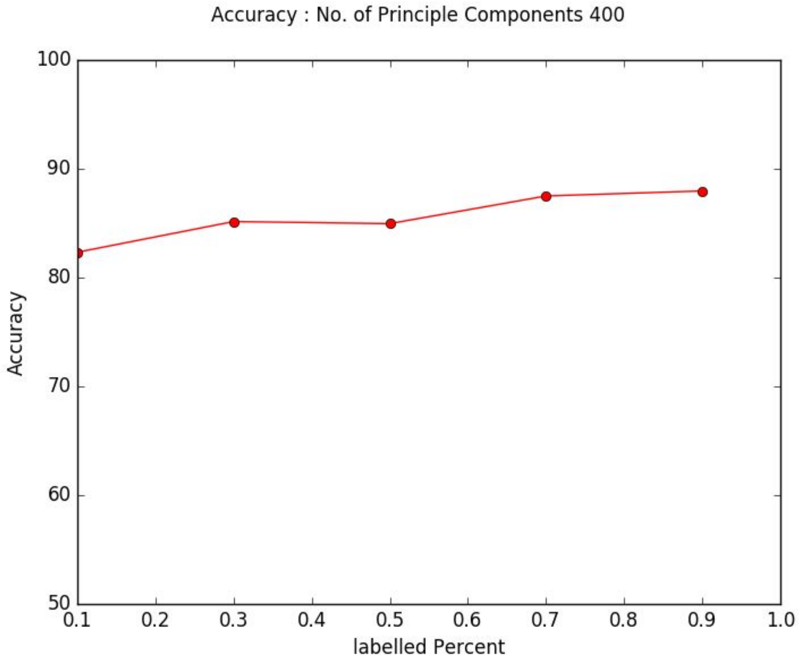
Top ‘N’ Principal component extraction

Train test splitting for K fold cross validation

Dumping text files

Experimental Results





Thank You !!!