Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition

Peg Howland and Haesun Park

Abhishek Kalokhe, Rajeshwari Devaramani

Georgia Institute of Technology

Background

- Discriminant Analysis has always been the widely used method for extracting the features from the data while preserving the cluster separability.
- It is commonly defined as an optimization problem involving covariance matrices that represent the scatter within and between clusters.

Background

Limitations Classical Discriminant Analysis methods had a requirement of one of the covariance matrices being nonsingular.

This limited it's application to datasets with certain relative dimensions.

Objectives The objectives of the paper address this problem and propose a general method eliminating the classical limitations.

Introduction

Introduction

- The goal of the paper is to combine the features of the original data in a way that maintains the cluster structure of the data.
- **Assumption:** The data is clustered.
- **What we want to achieve?**

$$G^T : a \in \mathbb{R}^{m \times 1} \rightarrow y \in \mathbb{R}^{l \times 1}$$

Introduction

- **Dataset Used:** Department of Justice 2009-2018 Press Releases
- Used the TfIdf Vectorizer which converts a collection of raw documents to a matrix of TF-IDF features.
- We have 300 documents (samples) with:
  - 1000 features for undersampled case
  - 50 features for oversampled case
- **Clusters:**

```
Cluster 0
natural,emissions,oil,environment,clean,water,settlement,environmental,air,epa

Cluster 1
elections,bailout,voters,activities,observers,monitor,county,election,rights,voting

Cluster 2
taxes,refunds,trial,indictment,prison,false,income,returns,irs,tax
```

https://www.kaggle.com/datasets/jbencina/department-of-justice-20092018-press-releases/data

Introduction

- We represent the vectorized dataset as matrix A:
  $A = (A_1, A_2, ..., A_k)$ where $A_i \in \mathbb{R}^{m \times n_i}$ and $\sum_{i=1}^{k} n_i = n$

  Here, the data vectors $a_1, a_2, ..., a_n$ are the columns of matrix A.
- Let N_i denote the set of column indices that belong to cluster i. The centroid $c^{(i)}$ is computed by taking the average of the columns in cluster.

$$c^{(i)} = \frac{1}{n_i} \sum_{j \in N_i} a_j$$

and the global centroid is defined as,

$$c = \frac{1}{n} \sum_{j=1}^{n} a_j$$

Introduction

- We define scatter matrix S_W, S_B and S_M as:

$$S_W = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})(a_j - c^{(i)})^T$$

$$S_B = \sum_{i=1}^{k} \sum_{j \in N_i} (c^{(i)} - c)(c^{(i)} - c)^T$$

$$= \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T$$

$$S_W = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})(a_j - c^{(i)})^T$$

The relation between them is defined as $S_M = S_B + S_W$.
As we apply G^T to the matrix A, it transforms the scatter matrices to $l \times l$ matrices,

$$S_W^Y = G^T S_W G, S_B^Y = G^T S_B G, S_M^Y = G^T S_M G,$$

Cluster Quality

- When cluster quality is high, each cluster is tightly grouped, but well separated from the other clusters.
- $$\mathrm{trace}(S_W) = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c(i))^T (a_j - c(i)) = \sum_{i=1}^{k} \sum_{j \in N_i} \|a_j - c(i)\|^2$$

  measures the closeness of the columns within the clusters
- $$\mathrm{trace}(S_B) = \sum_{i=1}^{k} \sum_{j \in N_i} (c(i) - c)^T (c(i) - c) = \sum_{i=1}^{k} \sum_{j \in N_i} \|c(i) - c\|^2$$

  measures the separation between clusters
- **Optimal transformation:** maximize $\mathrm{trace}(S_Y^B)$ and minimize $\mathrm{trace}(S_Y^W)$.
- $$\max_{G} \mathrm{trace}((G^T S_2 G)^{-1} (G^T S_1 G))$$

Generalized Singular Value Decomposition

Generalized Singular Value Decomposition

- **Van-Loan**
  Suppose two matrices $K_A \in \mathbb{R}^{p\times m}$ with $p \geq m$ and $K_B \in \mathbb{R}^{n\times m}$ are given. Then, there exist orthogonal matrices $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{n\times n}$ and a nonsingular matrix $X \in \mathbb{R}^{m\times m}$ such that

$$U^T K_A X = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$$

$$V^T K_B X = \mathrm{diag}(\beta_1, \ldots, \beta_q)$$

  where $q = \min(n, m)$, $\alpha_i \geq 0$ for $1 \leq i \leq m$, and $\beta_i \geq 0$ for $1 \leq i \leq q$.

Generalized Singular Value Decomposition

- **Paige and Saunders**
  Given $K_A \in \mathbb{R}^{p\times m}$ and $K_B \in \mathbb{R}^{n\times m}$, there exist orthogonal matrices $U \in \mathbb{R}^{p\times p}$, $V \in \mathbb{R}^{n\times n}$, $W \in \mathbb{R}^{t\times t}$, and $Q \in \mathbb{R}^{m\times m}$ such that

$$U^T K_A Q = \Sigma_A (W^T R\,,\ 0)$$

$$V^T H^T w Q = \Sigma_B (W^T R\,,\ 0)$$

  where $K = \begin{pmatrix} K_A \\ K_B \end{pmatrix}$ and $t = \text{rank}(K)$, $R \in \mathbb{R}^{t\times t}$ is nonsingular with singular values equal to the nonzero singular values of K.

Generalized Singular Value Decomposition

- Relating to Van Loan's:

$$U^T K_A X = (\Sigma_A, 0) \quad \text{and} \quad V^T K_B X = (\Sigma_B, 0),$$

where $X_{m\times m} = Q \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}$.

Therefore,

$$K_A^T K_A = X^{-T} \begin{pmatrix} \Sigma_A^T \Sigma_A & 0 \\ 0 & 0 \end{pmatrix} X^{-1}$$

and

$$K_B^T K_B = X^{-T} \begin{pmatrix} \Sigma_B^T \Sigma_B & 0 \\ 0 & 0 \end{pmatrix} X^{-1},$$

and

$$\beta_i^2 K_A^T K_A x_i = \alpha_i^2 K_B^T K_B x_i \quad \text{for} \quad 1 \le i \le t.$$

Generalization of Linear Discriminant Analysis

Optimization of J_1 = trace($S_2^{-1}S_1$) Criteria

$$\text{Optimize } J_1(G) = \text{trace}((G^T S_2 G)^{-1}(G^T S_1 G))$$

over G, where S_1 and S_2 are chosen from S_W, S_B and S_M.Assume S_2 to be nonsingular, it is symmetric positive definite.There exists a nonsingular matrix $X \in \mathbb{R}^{m \times m}$

$$X^T S_1 X = \Lambda = \text{diag}(\lambda_1 ... \lambda_m) \text{ and } X^T S_2 X = I_m$$

(Symmetric Definite Generalized Eigenvalue Problem)Letting x_i denote the ith column of X, we have

$$S_1 x_i = \lambda_i S_2 x_i$$

which means λ_i and x_i are an eigenvalue-eigenvector pair of $S_2^{-1}S_1$.$\lambda_i \geq 0$ for $1 \leq i \leq m$ (S_1 is positive semidefinite)Largest $q = \text{rank}(S_1)$ λ_is are non-zero.

Optimization of $J_1 = \text{trace}(S_2^{-1}S_1)$ Criteria

$$J_1(G) = \text{trace}((\tilde{G}^T\tilde{G})^{-1}\tilde{G}^T\Lambda\tilde{G})$$

where $\tilde{G} = X^{-1}G$. $\tilde{G}$ has full rank provided G does, so we can write $\tilde{G} = QR$, where $Q \in \mathbb{R}^{m \times l}$ has orthonormal columns and R is nonsingular. We get,

$$J_1(G) = \text{trace}(Q^T\Lambda Q)$$

Once we have simultaneously diagonalized S_1 and S_2, the maximization of $J_1(G)$ depends only on an orthonormal basis for range($X^{-1}G$), i.e,

$$\max_{G} J_1(G) = \max_{Q^TQ=I_l} \text{trace}(Q^T\Lambda Q)$$

$$\leq \lambda_1 + ... + \lambda_q$$

$$= \text{trace}(S_2^{-1}S_1)$$

Optimization of J_1 = trace($S_2^{-1}S_1$) Criteria

- For any l satisfying $l \geq q$, this upper bound on $J_1(G)$ is achieved for

$$Q = \begin{pmatrix} I_l \\ 0 \end{pmatrix} \text{ or } G = X \begin{pmatrix} I_l \\ 0 \end{pmatrix} R$$

- Tranformation G is not unique as J_1 satisfies invariance property $J_1(G) = J_1(GW)$ for any nonsingular matrix $W \in \mathbb{R}^{l \times l}$.
- Hence, maximum $J_1(G)$ is also achieved for

$$G = X \begin{pmatrix} I_l \\ 0 \end{pmatrix}$$

This means that, for $l \geq \text{rank}(S_1)$,

$$\text{trace}((G^T S_2 G)^{-1} G^T S_1 G) = \text{trace}(S_2^{-1} S_1)$$

whenever $G \in \mathbb{R}^{m \times l}$ consists of l eigenvectors of $S_2^{-1}S_1$ corresponding to the l largest eigenvalues.

Optimization of J_1 = trace($S_2^{-1}S_1$) Criteria

- According to our partitioning of A into k clusters, we define $m \times n$ matrices,

$$H_W = (A_1 - c^{(1)}e^{(1)^T}, A_2 - c^{(2)}e^{(2)^T}, ..., A_k - c^{(k)}e^{(k)^T})$$

$$H_B = ((c^{(1)} - c)e^{(1)^T}, (c^{(2)} - c)e^{(2)^T}, ..., (c^{(k)} - c)e^{(k)^T})$$

$$H_M = (a_1 - c, ..., a_n - c) = A - ce^T = H_W + H_B$$

where $e^{(i)} = (1, ..., 1)^T \in \mathbb{R}^{n_i \times 1}$ and $e = (1, ..., 1)^T \in \mathbb{R}^{n \times 1}$

- Scatter matrices are expressed as

$$S_W = H_W H_W^T, S_B = H_B H_B^T, S_M = H_M H_M^T$$

- J_1 cannot be applied when the number of available data vectors n is smaller than the dimension m of the data.
- We generalize by expressing λ_i as α_i^2/β_i^2 in

$$S_1 x_i = \lambda_i S_2 x_i$$

to,

$$\beta^2 S_i x_i = \alpha_i^2 S_2 x_i$$

Generalization of J_1 = trace($S_2^{-1}S_1$) Criteria for Singular S_2

- **Case 1:**

$$(S_1, S_2) = (S_B, S_W).$$

To approximate G that satisfies both

$$\max_G trace(G^T S_B G) \quad \text{and} \quad \min_G trace(G^T S_W G),$$

- For nonsingular S_W, the generalized singular vectors are eigenvectors of $S_W^{-1}S_B$, so we choose the $x_i s$ which correspond to the k - 1 largest $\lambda_i s, where \lambda_i = \alpha_i^2/\beta_i^2$
- When m > n, the scatter matrix S_W is singular. Hence, the eigenvectors of $S_W^{-1}S_B$ are undefined, and classical discriminant analysis fails
- If a generalized singular vector x_i lies in the null space of S_W From equation, we see that either x_i also lies in the null space of S_B, or the corresponding β_i equals zero.

Equivalence of J_1 = trace$(S_2^{-1}S_1)$ Criteria for various S_2 and S_1

- Case when $(S_1, S_2) = (S_M, S_W)$ according to our previous analysis we would have to include rank(S_M) columns of X in G, which is not less than or equal to $k-1$. However,

$$S_M x_i = \lambda_i S_W x_i$$

can be written as

$$S_B x_i = (\lambda_i - 1) S_W x_i, \text{ where } \lambda_i \geq 1 \text{ for } 1 \leq i \leq m$$

- In this case, the eigenvector matrix is the same as for the case of $(S_1, S_2) = (S_B, S_W)$, but the eigenvalue matrix is $\Lambda - I$.
- Same permutation will put the $\Lambda - I$ in nonincreasing order as was used for Λ, and x_i corresponds to the ith largest eigenvalue of $S_W^{-1}S_B$, therefore, for nonsingular S_W, the solution is same as for $(S_1, S_2) = (S_B, S_W)$.

Alternative Approches

Alternative Approches

- Orthogonal Centroid
- Simpler criteria for preserving the cluster structure.
- Involve only one of the scatter matrices, min trace($G^T S_W G$) or max trace ($G^T S_B G$).
- min trace($G^T S_W G$) is meaningless as the optimum always reduces the dimension to one, even with the restriction of G having orthomormal columns.
- With same restriction, maximization of trace ($G^T S_B G$) produces solution equivalent to orthogonal centroid method.
- Let $J_2(G) = \textbf{trace}\left(G^T S_B G\right)$ and $G \in \mathbb{R}^{m \times l}$ has orthonormal columns, then there exists $\hat{G} \in \mathbb{R}^{m \times (m-l)}$ such that $\left[G, \hat{G}\right]$ is an orthogonal matrix.

Alternative Approches

- **Orthogonal Centroid**
- Since S_B is positive semidefinite,

$$\mathrm{trace}\left(G^T S_B G\right) \leq \mathrm{trace}\left(G^T S_B G\right) + \mathrm{trace}\left(\hat{G}^T S_B \hat{G}\right) = \mathrm{trace}\left(S_B\right).$$

- If SVD of H_B is given by $H_B = U\Sigma V^T$, then $S_B U = U\Sigma\Sigma^T$.
- Columns of U form an orthonormal set of eigenvectors of S_B corresponding to the nonincreasing eigenvalues σ_i on the diagonal of $\Lambda = \Sigma\Sigma^T$
- For $q = \mathrm{rank}\left(S_B\right)$, if we let U_q denote the first q columns of U and $\Lambda_q = \mathrm{diag}\left(\sigma_1, \ldots, \sigma_q\right)$, we have

$$J_2\left(U_q\right) = \mathrm{trace}\left(U_q^T S_B U_q\right) = \mathrm{trace}\left(U_q^T U_q \Sigma_q\right) = \lambda_1 + \cdots + \lambda_q = \mathrm{trace}\left(S_B\right)$$

- We can see that trace(S_B) is preserved when we take U_q as G.

Alternative Approches

- **Orthogonal Centroid**
- We define a centroid matrix $C = (c^{(1)}, c^{(2)}, \ldots, c^{(k)})$
- C has reduced QR decomposition $C = Q_k R$, where $Q_k \in \mathbb{R}^{m\times k}$ has orthonormal columns and $R \in \mathbb{R}^{k\times k}$.
- Let x be an eigenvector corresponding to nonzero eigenvalue λ, then

$$S_B x = \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T x = \lambda x$$

- which means $x \in \mathrm{span}\{c^{(i)} | 1 \le i \le k\}$
- Hence, we have $\mathrm{range}(U_q) \subseteq \mathrm{range}(C) \subseteq \mathrm{range}(Q_k)$, which implies that $U_q = Q_k W$ for some matrix $W \in \mathbb{R}^{k\times q}$ with orthonormal columns.

Alternative Approches

- **Orthogonal Centroid**
- We get,

$$J_2(U_q) = \mathrm{trace}(W^T Q_k^T S_B Q_k W) \leq \mathrm{trace}(Q_k^T S_B Q_k) = J_2(Q_k)$$

Hence, $J_2(Q_k) = \mathrm{trace}(S_B)$ and therefore by computing reduced QR of the centroid matrix, we obtain a solution that maximizes the $\mathrm{trace}(G^T S_B G)$ over all G with orthonormal columns.

Alternative Approches

- **Two-Stage Approach**
- Another approach for dealing with the singularity of S_W when $m > n$.
- As the name suggests, this approach works in two stages.

  First Using LSI/SVD, reduce the dimension of the data enough so that the new S_W is nonsingular.

  Second Perform classical LDA.
- Truncated SVD is used to find rank-l approximation of A.
- If $l \leq \mathrm{rank}(A)$, then

$$A \approx U_l \Sigma_l V_l^T$$

- LSI/SVD uses $\Sigma_l V_l^T$ as the reduced dimensional representation of A or equivalently computes the l-dimensional representation of $a \in \mathbb{R}^{m \times 1}$ as $y = U_l^T a$.

Algorithms

Algorithm 1: LDA/GSVD

Algorithm 1 LDA/GSVD

Given a data matrix $A \in \mathbb{R}^{m\times n}$ with k clusters and an input vector $a \in \mathbb{R}^{m\times 1}$, compute the matrix $G \in \mathbb{R}^{m\times(k-1)}$ which preserves the cluster structure in the reduced dimensional space, using

$$J_1(G) = \text{trace}((G^T S_W G)^{-1} G^T S_B G).$$

Also compute the $k-1$ dimensional representation y of a.

1) Compute H_B and H_W from A according to

$$H_B = (\sqrt{n_1}(c^{(1)} - c), \sqrt{n_2}(c^{(2)} - c), \ldots, \sqrt{n_k}(c^{(k)} - c))$$

and (11), respectively. (Using this equivalent but $m \times k$ form of H_B reduces complexity.)

2) Compute the complete orthogonal decomposition

$$P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}, \text{ where } K = \begin{pmatrix} H_B^T \\ H_W^T \end{pmatrix} \in \mathbb{R}^{(k+n)\times m}$$

3) Let $t = \text{rank}(K)$.

4) Compute W from the SVD of $P(1:k, 1:t)$, which is

$$U^T P(1:k, 1:t) W = \Sigma_A.$$

5) Compute the first $k-1$ columns of $X = Q\begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}$, and assign them to G.

6) $y = G^T a$

Algorithm 2: Orthogonal Centroid

Algorithm 2 Orthogonal Centroid

Given a data matrix $A \in \mathbb{R}^{m \times n}$ with k clusters and an input vector $a \in \mathbb{R}^{m \times 1}$, compute a k-dimensional representation y of a.

1) Compute the centroid $c^{(i)}$ of the ith cluster, $1 \leq i \leq k$.
2) Set $C = (c^{(1)}, c^{(2)}, \ldots, c^{(k)})$.
3) Compute the matrix Q_k in the reduced QR decomposition $C = Q_k R$.
4) $y = Q_k^T a$.

RESULTS

RESULTS

TABLE 1
Traces and Misclassification Rates (in Percent)
with L_2 Norm Similarity

Method	Full	trace($S_W^{-1}S_B$)	trace($S_W^{-1}S_M$)	
Dim	150×2000	6×2000	6×2000	7×2000
trace(S_W)	299700	1.97	1.48	1.98
trace(S_B)	22925	4.03	3.04	3.04
trace(S_M)	322630	6.00	4.52	5.02
trace($S_W^{-1}S_B$)	12.6	12.6	12.6	12.6
trace($S_W^{-1}S_M$)	162.6	18.6	18.6	19.6
centroid	2.6 %	2.2 %	2.0 %	2.0 %
5nn	18.7 %	2.2 %	2.2 %	2.4 %
15nn	10.1 %	1.8 %	1.9 %	2.1 %

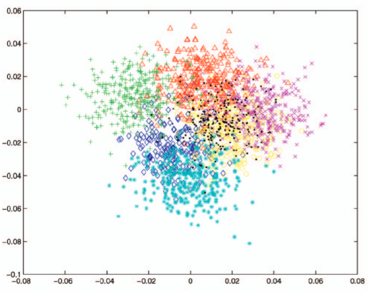
- Here, they use clustered data that are artificially generated by an algorithm. The data consist of 2,000 vectors in a space of dimension 150, with k=7 clusters.

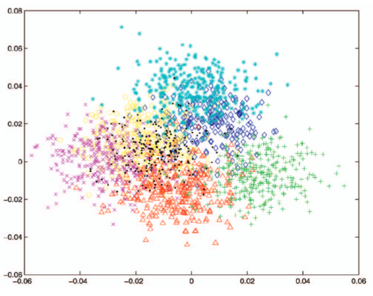
RESULTS

- LDA/GSVD reduces the dimension from 150 to k−1=6
- Comparison of LDA/GSVD criterion, $J_1 =$ $\text{trace}(S_W^{-1}S_B)$ and $\mathrm{trace}(S_W^{-1}S_M)$
- The trace values confirm our theoretical findings, namely, that the generalized eigenvectors that optimize the alternative J1 also optimize LDA/GSVD's J1, and including an additional eigenvector increases $\text{trace}(S_W^{-1}S_M)$ by one.
- reports misclassification rates for a centroid-based classification method and the k nearest neighbor show no advantage of using S_M over S_B
- These results bolster choice of $J_1 = \text{trace}(S_W^{-1}S_B)$ in the LDA/GSVD algorithm since it limits the GSVD computation to a composite matrix with k + n rows, rather than one with 2n row

Discriminatory Power of J_1

- We apply it to the same 2,000 data vectors, this time we reduce the dimension from 150 to two.
- Even though the optimal reduced dimension is six, $J_1 = \mathrm{trace}(S_W^{-1} S_B)$does surprisingly well at discriminating among seven classes. $J_1 = \mathrm{trace}(S_W^{-1} S_M)$ also does equally well





Conclusions

- Experimental results verify that the J_1 criterion, when applicable, effectively optimizes classification in the reduced dimensional space, while our LDA/GSVD extends the applicability to cases that classical discriminant analysis cannot handle.
- In addition, our LDA/GSVD algorithm never explicitly forms the scatter matrices, which results in two advantages.
- we avoid the numerical problems inherent in forming cross-product matrices.
- we reduce the storage requirements considerably.