Similarity Based Constraint Score

Mid-term Research Project Presentation

Gadegbeku Fabio

January 24, 2024

Table of Contents

Introduction

Problem : Too many Features

- Curse of dimensionality
- Unnecessarily high computational cost

Solution : Feature Selection

- Tackles the curse of dimensionality
- Removes irrelevant or redudant features

Introduction

Classification Problems

- We can define **must link** and **cannot link** constraints
  - must link : When two samples have the same class
  - cannot link : When two samples have different classes
- **Constraint Scores** to evaluate how well each feature respects the constraints

Introduction

Constraint Scores

- Typically compute distances between the samples in the original feature space
- Still suffer from the curse of dimensionality

Solution : Similarity Based Constraint Score (SBCS)

- Evaluate a whole subset of features at once
- Calculates distances in a lower dimensional space

Introduction

Goals

- Implement the SBCS as described by [1]
- Compare it to other constraint scores on different datasets on multiple criteria
- Improve the SBCS by using constraints directly instead of available labels to generate the constraints

Laplacian Score

Our Data

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1r} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ir} & \cdots & x_{id} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nr} & \cdots & x_{nd} \end{bmatrix}$$

A sample of our Data

$$x_i = (x_{i1}, \ldots, x_{ir}, \ldots, x_{id})^T \in \mathbb{R}^d$$

A feature vector

$$f_r = (x_{1r}, \ldots, x_{ir}, \ldots, x_{nr})^T \in \mathbb{R}^n$$

Laplacian Score

Similarity Matrix

$$W = \begin{bmatrix} 1 & w_{12} & \dots & w_{1n} \\ w_{21} & 1 & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & 1 \end{bmatrix}$$

Similarity between two samples

$$w_{ij} = S(x_i, x_j)$$

For example :

$$S(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$$

Laplacian Score

Degree Matrix

$$D = \begin{bmatrix} d_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_{nn} \end{bmatrix}$$

Where :

$$d_{ii} = \sum_{j=1}^{n} w_{ij}$$

Laplacian Score

Laplacian Matrix

$$L = D - W$$

Laplacian Score

$$L_r = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ir} - x_{jr})^2 s_{ij}}{\sum_{i=1}^{n} (x_{ir} - \bar{f}_r) p_i}$$

Where :

$$pi = \frac{d_i}{\sum_{k=1}^{n} d_k}$$

And we have :

$$L_r = \frac{{f_r}^T L f_r}{{f_r}^T D f_r}$$

Constraint Score 1

Constraints

$$\mathcal{M} = \{(x_i, x_j) \in X \times X \,|\, \text{such that } x_i \text{ and } x_j \text{ are together}\}$$

$$\mathcal{C} = \{(x_i, x_j) \in X \times X \,|\, \text{such that } x_i \text{ and } x_j \text{ are not together}\}$$

We can define :

$$w_{ij}^{\mathcal{M}} = \begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{M} \text{ or } (x_j, x_i) \in \mathcal{M} \\ 0 & \text{else} \end{cases}$$

$$w_{ij}^{\mathcal{C}} = \begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{C} \text{ or } (x_j, x_i) \in \mathcal{C} \\ 0 & \text{else} \end{cases}$$

Constraint Score 1

Minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir} - x_{jr})^2 w_{ij}^{\mathcal{M}}$$

Maximize

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir} - x_{jr})^2 w_{ij}^{\mathcal{C}}$$

Constraint Score 1

New "Laplacian" Matrix

$$L^{\mathcal{M}} = D^{\mathcal{M}} - W^{\mathcal{M}} \quad \text{and} \quad L^{\mathcal{C}} = D^{\mathcal{C}} - C^{\mathcal{C}}$$

Where :

$$D_{ii}^{\mathcal{M}} = \sum_{j=1}^{n} w_{ij}^{\mathcal{M}} \quad D_{ii}^{\mathcal{C}} = \sum_{j=1}^{n} w_{ij}^{\mathcal{C}}$$

Constraint Score 1

$$SC_r^1 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ir} - x_{jr})^2 w_{ij}^{\mathcal{M}}}{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ir} - x_{jr})^2 w_{ij}^{\mathcal{C}}} = \frac{f_r^T L^{\mathcal{M}} f_r^T}{f_r^T L^{\mathcal{C}} f_r^T}$$

Constraint Score 4

SC4

$$SC_r^4 = \frac{{f_r}^T L f_r}{{f_r}^T D f_r} \cdot \frac{{f_r}^T L^{\mathcal{M}} f_r}{{f_r}^T L^{\mathcal{C}} f_r} = SL_r \cdot SC_r^1$$

Similarity Based Constraint Score

Subset of Features

$$F_m = \{f_1, \ldots, f_m\}$$

Supervised Learning

$$\hat{w}_{ij}^S = \begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{M} \\ 0 & \text{if } (x_i, x_j) \in \mathcal{C} \\ w_{ij}(F_m) & \text{otherwise} \end{cases}$$

Similarity Based Constraint Score

Semi Supervised Learning

$$w_{ij}^{SS} = \begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{M}^{SS} \\ 0 & \text{otherwise} \end{cases}$$

Where :

$$\mathcal{M}^{SS} = \{(x_i, x_j) \in X^2 \mid \exists l = 1, \ldots, k \text{ such that } NP(x_i) \in X^l \text{ and } NP(x_j) \in X^l\}$$

Similarity Based Constraint Score

Finally the score

$$\varepsilon^*(F_m) = \sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij}(F_m) - \hat{w}_{ij}^*)^2 = \|W(F_m) - \hat{W}^*\|_2 \text{ Where } * = \text{S or SS}$$

Similarity Based Constraint Score

Algorithm 1: Feature Selection Procedure

Input: Set of d features $F_d = \{f_1, \ldots, f_r, \ldots, f_d\}$.
Output: Subset of $\hat{m}$ relevant features $F_{\hat{m}}$.
$F_0 \leftarrow \{\emptyset\}$
for $m = 1$ *to* d **do**

Select the most relevant feature f_r^+:

$$f_r^+ = \arg \min_{f_r \in F_d \setminus F_{m-1}} (\varepsilon^*(F_{m-1} \cup \{f_r\}))$$

Update $F_m \leftarrow F_{m-1} \cup \{f_r^+\}$

Select the number $\hat{m}$ of features such that:

$$\hat{m} = \arg \min_{m=1,2,\ldots,d} (\varepsilon^*(F_m))$$

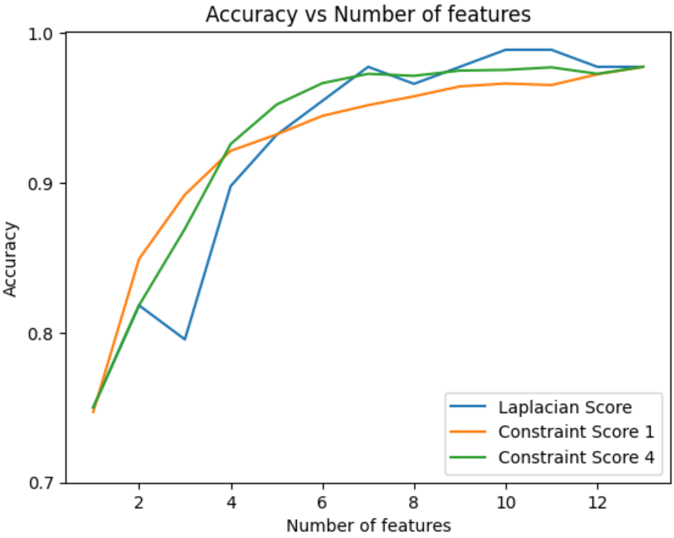
Output: $F_{\hat{m}}$

Results

Wine Database

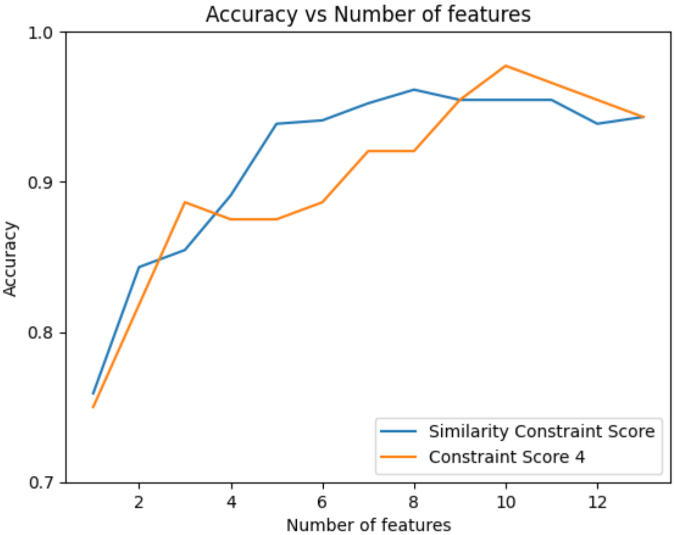
- 178 samples characterized by 13 features (n=178, d=13)
- 3 classes (k=3)
  - 59 class 1
  - 71 class 2
  - 48 class 3
- We select 30, 36, and 24 instances from each class to constitute the training set.
- 1-NN classifier to measure accuracy
- 9 labels avaiable (3 prototypes per class)

Results



Accuracy vs Number of features
1.0
0.9
0.8
0.7
Accuracy
2
4
6
8
10
12
Number of features
Laplacian Score
Constraint Score 1
Constraint Score 4

Results



Accuracy vs Number of features
1.0
0.9
0.8
0.7
Accuracy
2
4
6
8
10
12
Number of features
Similarity Constraint Score
Constraint Score 4

[1] Abderezak Salmi, Kamal Hammouche, and Ludovic Macaire. "Similarity-Based Constraint Score for Feature Selection". In: *Knowledge-Based Systems* 209 (Dec. 2020), p. 106429. ISSN: 09507051. DOI: 10.1016/j.knosys.2020.106429. (Visited on 09/23/2023).