Similarity Based Constraint Score For Feature Selection

Research Project Report

Gadegbeku Fabio

Supervised by:
Dr. Ludovic Macaire

April 14, 2024

Abstract

To avoid high dimensional feature spaces in machine learning tasks, feature selection methods are used to select a subset containing the most relevant features. In classification problems several scores have been proposed to evaluate the relevance of features, among them are constraint scores that use prior knowledge to define constraints that the features must respect to be qualified as relevant. These scores typically compute distances between samples in the original feature space, which still suffers from the curse of dimensionality, also they ignore the correlation between features. In this paper we propose a new constraint score based on the similarity between samples in a lower dimensional space that takes into account the correlation between features. We evaluate the performance of this score on benchmark datasets and compare it to other state of the art constraint scores.

Contents

1 Introduction

In Machine Learning having too many features is counter productive, this is called the curse of dimensionality. To avoid this phenomenon there exists feature selection methods that evaluate the revelance of the features. These feature selection techniques fall into three main categories: filter, wrapper, and embedded methods. Filter methods assess features without considering the classification algorithm, whereas wrapper methods utilize a classification algorithm to gauge feature relevance, lastly embedded methods incorporate feature selection directly into the learning algorithm. In our case we focus on filter methods for feature selection.

Filter methods can be further divided into supervised and unsupervised methods. Supervised methods use the labels to evaluate the features, whereas unsupervised methods do not use the labels. In this report we will focus on semi-supervised methods that can use prior knowledge provided by an expert on a subset of the data called the prototype set. This prior knowledge is used to define constraints that the features must respect. More precisely in classification problems we can use *must link*[1] and *cannot link*[2] constraints to define constraint scores to evaluate how well each feature respects the constraints. These constraint scores typically compute distances between the samples in the original feature space to evaluate them, so still suffer from the curse of dimensionality.

In this paper we will present and implement the Similarity Based Constraint Score (SBCS) described by [5]. That has the unique capibilities of evaluating a whole subset of features at once and calculating distances in a lower dimensional space. We will also implement implement other state of the art constraint scores used for feature selection and compare them on several performance metrics using benchmark datasets.

1.1 Definitions and Notations

In this section we will define the basic notations and concepts used in this paper. First of all, $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{n \times d}$ denotes the dataset : n training data samples defined in a d-dimensional feature space. where $x_i = [x_{i1}, x_{i2}, \ldots, x_{ir}, \ldots, x_{id}] \in \mathbb{R}^d$ is the i-th data sample of X, and x_{ir},$(r = 1, \ldots, d)$ is the r-th feature value of the i-th data sample. $F_d = \{f_1, f_2, \ldots, f_r, \ldots, f_d\}$ denote the set of d feature vectors of X, where $f_r = [x_{1r}, x_{2r}, \ldots, x_{nr}]^T \in \mathbb{R}^n$ is the r-th feature vector.

In general, prior knowledge is defined by an expert in the form of constraints, saying these two samples must be in the same class or these two samples must be in different classes. In our case we'll say we have a few labeled samples (protoypes) in each of the k classes and we'll deduce constraints from these labels.

2 Feature Scores

In this section we'll start by going more in depth into filter methods and how can they be used for feature selection in a (semi)-supervised and unsupervised context. Next we will present the different feature scores used in this paper. Starting with the state of the art constraint scores and then presenting the similarity based constraint score.

We seperated the scores into two sections : supervised and semi-supervised. Supervised scores are scores than can operate with only a few labels (prototype set) in X. The use only constraints defined by the expert. Semi-supervised scores are scores that can operate with only a few labels (prototype set) in X. They use constraints defined by the expert and also constraints defined by the algorithm. Making them partly unsupervised because they deduce new constraints from the prototype set.
In this sense the similarity based constraint score will have a supervised and a semi-supervised version.

NB : The names given to these scores are abritrary and may differ in other papers.

[1] When two samples have the same class
[2] When two samples have different classes

2.1 Filter Methods in Feature Selection

Consider this first instance :

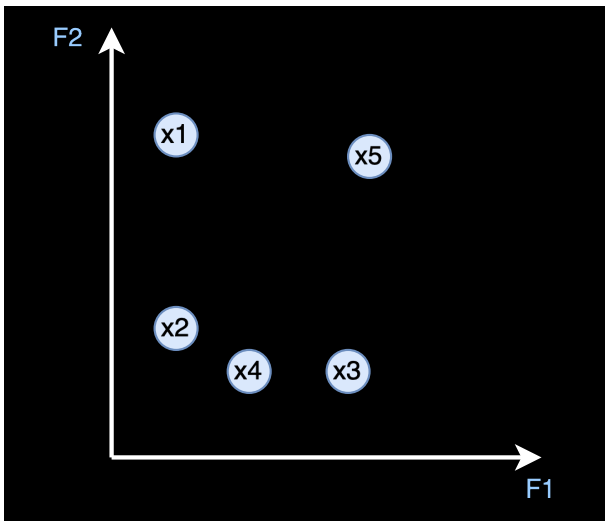


Figure 1: unsupervised feature selection

We have 5 samples represented by 2 features. We suppose we have no prior knowledge on the samples so no constraints are defined. We can use an unsupervised filter method to evaluate the relevance of the features. We can see that feature 2 seems to be more discrimating because samples far from each other in the original 2D feature space are far away if we project only on to F2. So an unsupervised filter method would most likely select feature 2 as the most relevant.

Consider this second instance :

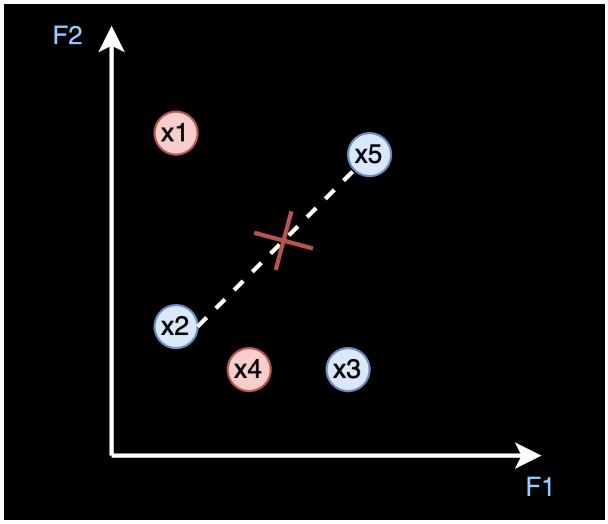


Figure 2: semi-supervised feature selection

Now we consider that we have prior knowledge on the samples. x_1 and x_4 are in the same class (RED) so we have a must link constraint between them. We also consider that we know that x_2 and x_5 are in different classes so we have a cannot link constraint between them.

In this new situation if we project onto F2 x_1 and x_4 are far away, whereas if we project onto F1 x_2 and x_5 are far away and x_1 and x_4 are close. So a semi-supervised filter method would select feature 1 as the most relevant.

2.2 Laplacian Score

In spectral graph theory our dataset X can be represented as an undirected weighted graph graph $G = (V, E, W)$ where the samples are the vertices V and the edges E are defined by the similarity matrix W. The similarity matrix W is a $n \times n$ matrix where w_{ij} is the similarity between samples x_i and x_j. In general we used the Gaussian similarity function :

$$w_{ij} = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}} \tag{1}$$

The Laplacian matrix is defined as :

$$L = D - W \tag{2}$$

where $D \in \mathbb{R}^{n \times n}$ is the diagonal degree matrix whose elements are $d_{ii} = \sum_{j=1}^{n} w_{ij}$.
With this information in an unsupervised framework He et al. [1] suppose that samples close to each other in the original feature space should be close to each other in the lower dimensional space.

This property can be measured by the Laplacian score of a feature f_r :

$$SL_r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir} - x_{jr})^2 s_{ij}}{\sum_{i=1}^{n}(x_{ir} - \bar{f}_r)p_i} \tag{3}$$

Where : $pi = \frac{d_i}{\sum_{k=1}^{n} d_k}$ can be seen as a probability density over our samples and $\bar{f}_r = \sum_{i=1}^{n} x_{ir} p_i$ is the ponderated mean of the feature vector f_r. And using the Laplacian matrix L we can rewrite this as :

$$L_r = \frac{f_r^T L f_r}{f_r^T D f_r} \tag{4}$$

A low score indicates that the feature respects the topological properties of the original feature space. SL_r operates in a completely unsupervised manner and there's no incoroparation of prior knowledge in the selection of features. This is not always relevant as seen in Figure 2.

2.3 Supervised Constraint Scores

To utilize constraint scores we need to define pairwise constraints between samples. We utilize the prototype set P to define these constraints, where P is a subset of X containing a few samples from each of the k classes. The constraints are defined as follows :

- Must Link : $(x_i, x_j) \in M$ if x_i and x_j are in the same class.
- Cannot Link : $(x_i, x_j) \in C$ if x_i and x_j are in different classes.

With these sets we can define two similarity matrices $W^M \in \mathbb{R}^{n \times n}$ and $W^C \in \mathbb{R}^{n \times n}$ defined by :

$$w_{ij}^M = \begin{cases} 1 & \text{if } (x_i, x_j) \in M \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$w_{ij}^C = \begin{cases} 1 & \text{if } (x_i, x_j) \in C \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

In this paper we consider two constraint scores by Zhang et al. [6], that operate in a supervised learning context. It uses the constraints defined by the expert to evaluate the features. While unconstrained samples are not taken into account. The first constraints scores of feature f_r are defined as follows :

$$SC_r^1 = \frac{\sum_{(x_i,x_j)\in M}(x_{ir} - x_{jr})^2}{\sum_{(x_i,x_j)\in C}(x_{ir} - x_{jr})^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir} - x_{jr})^2 w_{ij}^M}{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir} - x_{jr})^2 w_{ij}^C} \tag{7}$$

$$\begin{aligned} SC_r^2 &= \sum_{(x_i,x_j)\in M}(x_{ir} - x_{jr})^2 - \lambda \sum_{(x_i,x_j)\in C}(x_{ir} - x_{jr})^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir} - x_{jr})^2 w_{ij}^M - \lambda \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ir} - x_{jr})^2 w_{ij}^C \end{aligned} \tag{8}$$

Where λ is a parameter that controls the importance of the cannot link constraints. It is set to 1 in this original paper.

The goal of these scores is to maximize the distance between samples in the same class and minimize the distance between samples in different classes. To do so they evaluate how close the must link samples are and how far the cannot link samples are projected on feature f_r, a low score indicates that the feature is relevant. In practice if we wish to create a subset of size m we can sort the features by their scores and select the m best ones, i.e the ones with the lowest scores. Another option is to define a threshold and select the features with scores below this threshold.
For a more compact notation or easier implementation they can be defined using new laplacians matrices $L^M = D^M - W^M$ and $L^C = D^C - W^C$:

$$SC_r^1 = \frac{f_r^T L^M f_r}{f_r^T L^C f_r} \tag{9}$$

$$SC_r^2 = f_r^T L^M f_r - \lambda f_r^T L^C f_r \tag{10}$$

2.4 Semi-Supervised Constraint Scores

In this section we will present the semi-supervised versions of constraint scores. These scores operate in the same way with only a few labels (prototype set) in X. But they also deduce new constraints from the prototype set, and the local structure of the unlabeled samples.

The first semi-supervised constraint score that we will present is introduced by Zhao et al. [7], called the locality sensitive discriminant analysis score, refered to in this paper by SC_r^3. This score combines the similarity matrix W^C constructed from the cannot-link constraints and a new similarity matrix $W^{kn1} \in \mathbb{R}^{n \times n}$ which is constructed from the set of must-link constraints and unlabeled data samples as follows:

$$w_{ij}^{kn1} = \begin{cases} \gamma & \text{if } (x_i, x_j) \in M \\ 1 & \text{if } x_i \in X^U \text{ or } x_j \in X^U \text{ but } x_i \in KNN(x_j) \text{ or } x_j \in KNN(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where γ is a constant parameter. $KNN(x_i)$ is a set containing the K nearest neighbors of samplex_i In our experiments, γ is set to 100 and K is set to 5 as in [7].
The constraint score C_r^3 is defined as :

$$SC_r^3 = \frac{\sum_{i=1}^n \sum_{j=1}^n (x_{ir} - x_{jr})^2 w_{ij}^{kn1}}{\sum_{i=1}^n \sum_{j=1}^n (x_{ir} - x_{jr})^2 w_{ij}^C} = \frac{f_r^T L^{kn1} f_r}{f_r^T L^C f_r} \tag{12}$$

where $L^{kn1} = D^{kn1} - W^{kn1}$ is the Laplacian matrix of W^{kn1},and D^{kn1} the degree matrix computed from W^{kn1}.

The goal of this score is to maximize the distance between samples in the cannot-link set while also minimizing the distance between samples in the must-link set and the unlabeled samples that are close to each other.

The second semi-supervised constraint score that we will present is introduced by Kalakech et al. [2], refered to as SC_r^4 in this paper. This score is less sensitive to the constraints as it is a combination between the unsupervised Laplacian score SL_r and the supervised constraint score SC_r^1. The score is defined as :

$$SC_r^4 = \frac{f_r^T L f_r}{f_r^T D f_r} \cdot \frac{f_r^T L^M f_r}{f_r^T L^C f_r} = SL_r \cdot SC_r^1 \tag{13}$$

This score works in two parts, the laplacian score evaluates how well the topological properties of the original feature space and the constraint score evaluates how well the feature respects the constraints.

2.5 Similarity Based Constraint Score

As we have seen in the previous sections, existing constraint scores evaluate the the relevance of each feature independently and ranks them. In this section we will present the similarity based constraint

score introduced by Salmi et al. [5]. That evaluates the relevance of a subset of features at once. This score are based on the similarity between samples in a lower dimensional space and takes into account the correlation between features.

The approach used for this score is to compute the similarity matrix of a subset of m features $F_m = \{f_1, f_2, \ldots, f_m\}$ of X, $W(F_m) \in \mathbb{R}^{n \times n}$. We then compute a target similarity matrix $\hat{W}^*$ based on the constraints provided and measure the distance between $W(F_m)$ and $\hat{W}^*$. $*$ = supervised (S) or semi-supervised (SS).

The similarity matrix $W(F_m)$ is defined as :

$$w_{ij}(F_m) = \texttt{exp}\left(- \frac{\delta^2\big(x_i^{(m)}, x_j^{(m)}\big)}{2\sigma^2} \right) \tag{14}$$

Where $x_i^{(m)}$ is the vector of the i-th data point in the subset F_m.

The Supervised target similarity matrix $\hat{W^S}$ is defined as :

$$\hat{W}_{ij}^S = \begin{cases} 1 & \text{if } (x_i, x_j) \in M \\ 0 & \text{if } (x_i, x_j) \in C \end{cases} \tag{15}$$

The Semi-Supervised target similarity matrix $\hat{W^{SS}}$ is defined as :

$$\hat{W}_{ij}^{SS} = \begin{cases} 1 & \text{if } (x_i, x_j) \in M^{SS} \\ 0 & \text{if } (x_i, x_j) \in C \end{cases} \tag{16}$$

Where M^{SS} is a new set of must-link pairs deduced from prototype subsets X^l, $l = 1, \ldots, k$ and unlabeled data samples as follows:

$$M^{SS} = \Big\{ (x_i, x_j) \in X^2 | \exists l = 1, \ldots, k \quad \text{so that} \quad NP(x_i) \in X^l \quad \text{and} \quad NP(x_j) \in X^l \Big\} \tag{17}$$

Where $NP(x_i)$ is the nearest prototype of sample x_i in the original feature space. So in this case we also consider that we have a must link constraint between x_i and x_j if there nearest prototypes are in the same class, increasing the number of must link constraints.

The score is defined as :

$$\varepsilon^*(F_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} \Big(w_{ij}(F_m) - \hat{w}_{ij}^* \Big)^2 \tag{18}$$

Which can be written as the euclidean distance between two similarity matrices :

$$\varepsilon^*(F_m) = ||W(F_m) - \hat{W}^*||_2^2 \tag{19}$$

The goal of this score is to minimize the distance between the similarity matrix of the subset of features and the target similarity matrix $\hat{W}^*$. Since the score evaluates a whole subset of features at once and doesn't rank the features like previous scores we can use a greedy algorithm proposed by Salmi et al. [4] to select the a subset of features.

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: $F_{\hat{m}}$

3 Experimental Results

With the implementation of the different scores we can now evaluate their performance on benchmark datasets. We first compare the supervised scores (SC_r^1,SC_r^2,ε^S) and then the semi-supervised scores (SC_r^3,SC_r^4,ε^{SS}). Feature selection procedures were conducted on the training datasets across 100 runs. In each run, pairwise constraints were automatically generated. For each class , k prototype subsets X^l (with $|X^l| = 3$) were randomly selected from the standardized training dataset X. Subsequently, sets M, C, and M^{SS} of pairwise constraints.

3.1 Datasets

The experiments in this paper were conducted on 6 benchmark datasets from the UCI repository [3] Vehicles, Image Segmentation,Wisconsin Breast Cancer Diagnostic (WBCD),Ionosphere,Sonar and Libras Movement.

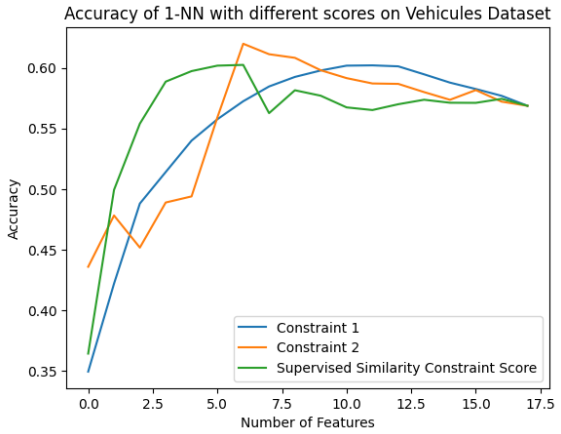
Dataset	Features d	# Training	# Test	Classes k	Prototypes $k * p$
Vehicles	18	846	846	4	12
Image Segmentation	19	210	210	7	21
WBCD	30	569	569	2	6
Ionosphere	34	351	351	2	6
Sonar	60	208	208	2	6
Libras Movement	90	360	360	15	45

Table 1: Datasets used in the experiments

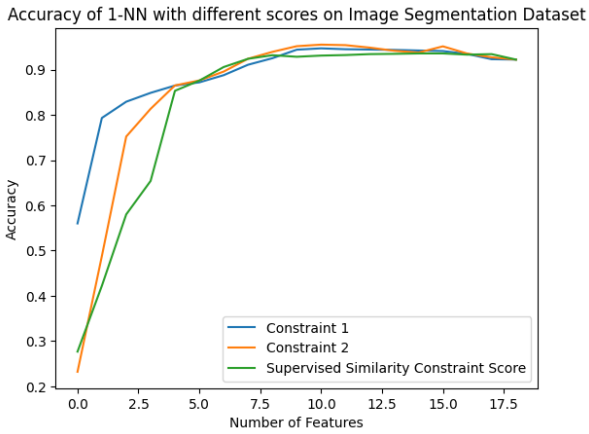
3.2 Accuracy

Our first metric was accuracy vs number of features. Figure 3 illustrates the average accuracy for ε^S, C^1, and C^2 obtained when the number of prototypes p is set to 3.

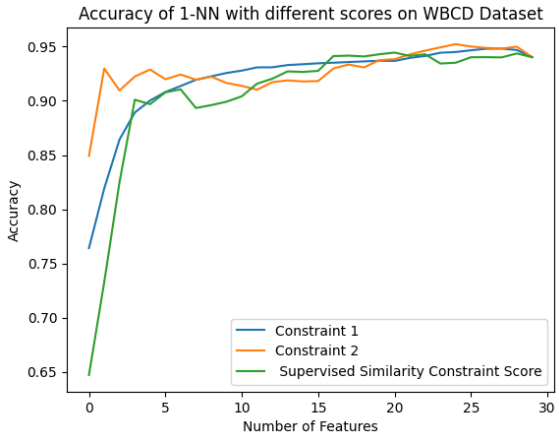
As we can see from the results the ε^S obtains correct results on all databases compared to SC_r^1 and SC_r^2, who vary a bit more performance wise. On the higher dimensional datasets like Libras Movement and Sonar ε^S outperforms the other scores.



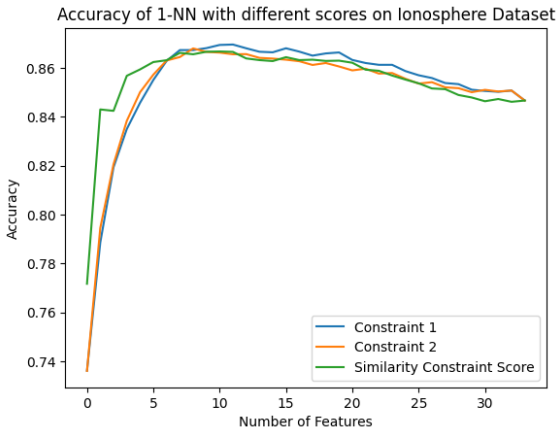
(a) Vehicles



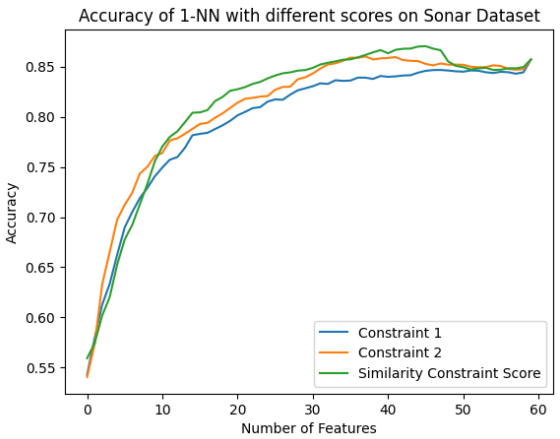
(b) Image Segmentation



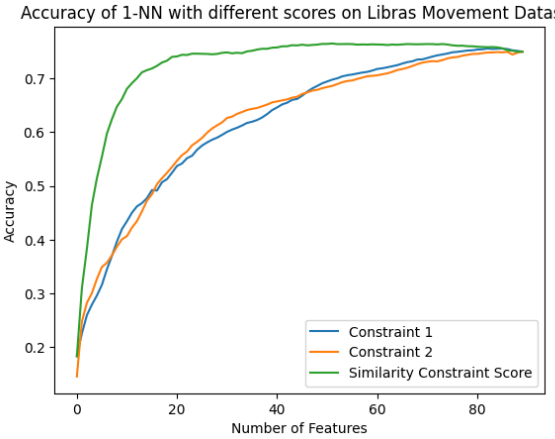
(c) WBCD



(d) Ionosphere



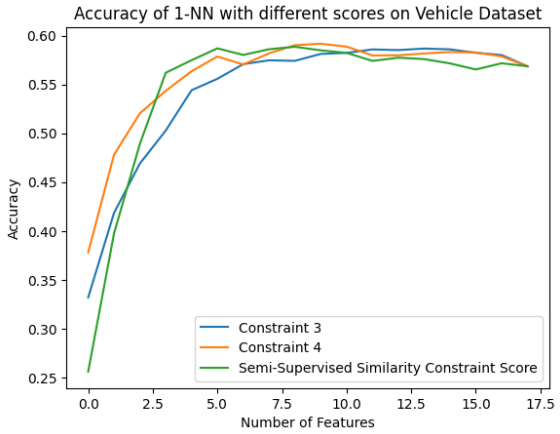
(e) Sonar



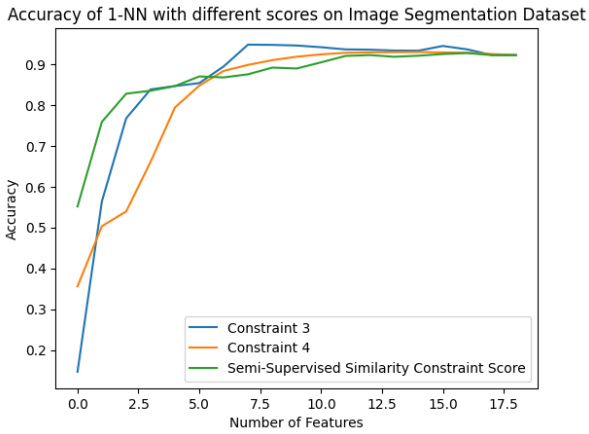
(f) Libras Movement

Figure 3: Accuracy versus number of selected features m by the supervised constraint scores on six benchmark databases.

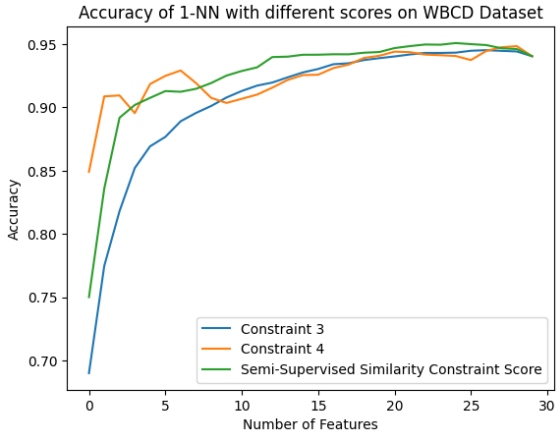
Next we have the results for the semi-supervised scores SC_r^3 and SC_r^4 and ε^{SS}. As seen in Figure 4 the ε^{SS} provides good results in general but is dependant on the accuracy of the new constraints deduced from the prototype set as shown in the plot for Sonar (e) we didn't get as good reults as ε^{S}.



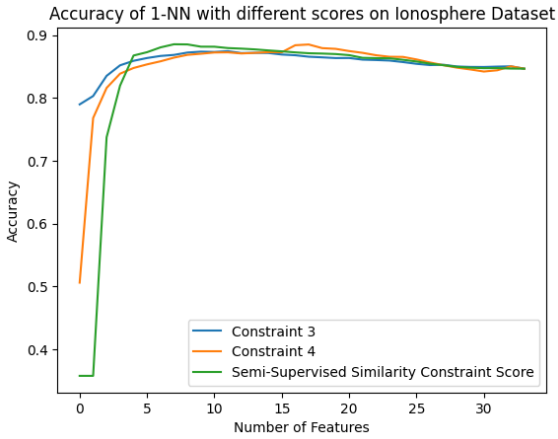
(a) Vehicles



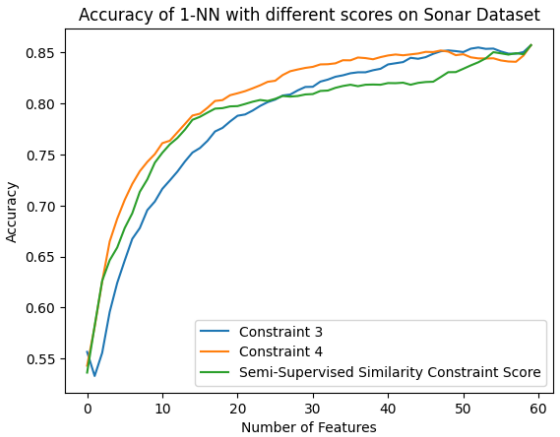
(b) Image Segmentation



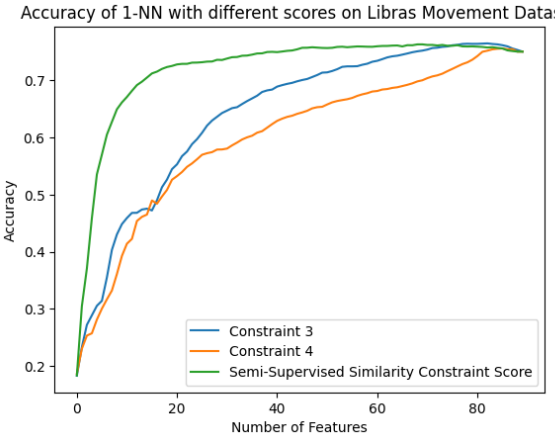
(c) WBCD



(d) Ionosphere



(e) Sonar



(f) Libras Movement

Figure 4: Accuracy versus number of selected features m by the semi-supervised constraint scores on six benchmark databases.

To go even further for the ε^{SS} we can see if the set of must link M^{SS} deduced from the prototype set are correct or not.

Table 2: Quality of deduced must links

Database	CR	COV	SMC
Vehicles	34.18%	38.34%	0.08%
Image Segmentation	52.72%	56.03%	0.03%
WBCD	77.66%	80.31%	0.01%
Ionosphere	61.12%	76.17%	0.02%
Sonar	51.81%	54.87%	0.08%
Libras Movement	42.49%	49.33%	1.5%

Where the CR is the ratio of the number of correct must link pairs to the total number of must link pairs generated, COV is the ratio of correct new must link pairs to the total number of possible must link pairs possible. And SMC is standard must link coverage ,the ratio of the number of must link pairs to the total number of possible must link pairs.

4 Conclusion

The paper introduces a novel constraint score for feature selection in supervised and semi-supervised settings. Unlike existing methods, this score score evaluates feature subsets together, identifying redundant or highly correletaed features. By assessing similarity among data samples within the feature subspace it avoids calculating distances in the original feature space, thus avoiding the curse of dimensionality.

In supervised learning, This score measures feature subsets constraint-preserving ability. In semi-supervised learning, new must-link constraints are derived from user-provided constraints and unlabeled data. Experiments on six benchmark databases show the performance of this score in comparison to existing constraint scores and how it clearly outperforms as the number of features increases.

References

[1] Xiaofei He, Deng Cai, and Partha Niyogi. "Laplacian Score for Feature Selection". In: ().

[2] Mariam Kalakech et al. "Constraint scores for semi-supervised feature selection: A comparative study". In: *Pattern Recognition Letters* 32.5 (2011), pp. 656–665.

[3] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: http://archive.ics.uci.edu/ml.

[4] Abderezak Salmi, Kamal Hammouche, and Ludovic Macaire. "Constrained Feature Selection for Semisupervised Color-Texture Image Segmentation Using Spectral Clustering". In: *Journal of Electronic Imaging* 30.01 (Feb. 2021). ISSN: 1017-9909. DOI: 10.1117/1.JEI.30.1.013014. (Visited on 09/25/2023).

[5] 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).

[6] Daoqiang Zhang, Songcan Chen, and Zhi-Hua Zhou. "Constraint Score: A New Filter Method for Feature Selection with Pairwise Constraints". In: *Pattern Recognition* 41.5 (May 2008), pp. 1440–1451. ISSN: 00313203. DOI: 10.1016/j.patcog.2007.10.009. (Visited on 09/24/2023).

[7] Jidong Zhao, Ke Lu, and Xiaofei He. "Locality Sensitive Semi-Supervised Feature Selection". In: *Neurocomputing* 71.10-12 (June 2008), pp. 1842–1849. ISSN: 09252312. DOI: 10.1016/j.neucom.2007.06.014. (Visited on 09/24/2023).