



**HAL**  
open science

## Constraint scores for semi-supervised feature selection: A comparative study

Mariam Kallakech, Philippe Biela, Ludovic Macaire, Denis Hamad

### ► To cite this version:

Mariam Kallakech, Philippe Biela, Ludovic Macaire, Denis Hamad. Constraint scores for semi-supervised feature selection: A comparative study. *Pattern Recognition Letters*, 2011, 32 (5), pp.656-665. 10.1016/j.patrec.2010.12.014 . hal-00732484

**HAL Id: hal-00732484**

**<https://hal.science/hal-00732484>**

Submitted on 14 Sep 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Constraint scores for semi-supervised feature selection: A comparative study

Mariam Kalakech<sup>a,b</sup>, Philippe Biela<sup>a,b</sup>, Ludovic Macaire<sup>b</sup>, Denis Hamad<sup>c</sup>

<sup>a</sup>*Hautes Etudes d'Ingénieur, 13 rue de Toul, F-59046, Lille, France*

<sup>b</sup>*LAGIS FRE CNRS 3303, Université de Lille 1 Nord de France, Bâtiment P2, Cité Scientifique, F-59655 Villeneuve d'Ascq, France*

<sup>c</sup>*LISIC, ULCO, 50, rue Ferdinand Buisson, F-62228 Calais, France*

---

## Abstract

Recent feature selection scores using pairwise constraints (must-link and cannot-link) have shown better performances than the unsupervised methods and comparable to the supervised ones. However, these scores use only the pairwise constraints and ignore the available information brought by the unlabeled data. Moreover, these constraint scores strongly depend on the given must-link and cannot-link subsets built by the user. In this paper, we address these problems and propose a new semi-supervised constraint score that uses both pairwise constraints and local properties of the unlabeled data. Experimental results show that this new score is less sensitive to the given constraints than the previous scores while providing similar performances. *Keywords:* Feature selection, Pairwise constraints, Kendall's coefficient, Constraint scores, Laplacian score, Fisher score

---

1 **1. Introduction**

2 In machine learning and pattern recognition applications, the process-  
3 ing of high dimensional data requires large computation time and capacity  
4 storage. Though, it leads to poor performances when the dimensionality to  
5 sample size ratio is high. To improve performances, the sample dimension-  
6 ality is reduced thanks to feature extraction or selection schemes (Liu and  
7 Motoda (1998); Yu and Liu (2003)). Let us notice that feature extraction  
8 transforms the original input space into a new low dimensional space by com-  
9 bining the initial features, while feature selection retains the most relevant  
10 ones in order to build a low dimensional feature space.

11 Data samples can be either unlabeled or labeled, leading to the development  
12 of unsupervised and supervised feature selection techniques. Unsupervised  
13 feature selection measures the feature capacity of keeping the intrinsic data  
14 structure in order to evaluate its relevance (Dy and Brodley (2004)). Super-  
15 vised feature selection consists in evaluating feature relevance by measuring  
16 the correlation between the feature and class labels (Yu and Liu (2004)).

17

18 Supervised feature selection requires sufficient labeled data samples in order

19 to provide a discriminating feature space. However, the sample labeling pro-  
20 cess by the human user is fastidious and expensive. That is the reason why in  
21 many real applications, we have huge unlabeled data and small labeled sam-  
22 ples. To deal with this "lack labeled-sample problem", recent semi-supervised  
23 feature selection schemes have been developed by Zhao et al. (Zhao and Liu  
24 (2007a); Zhao and Liu (2007b)). They propose a semi-supervised feature  
25 relevance criterion which takes into account both unlabeled data and labeled  
26 samples. Unfortunately, this score requires to define the classifier step in  
27 order to compare the initial labels of the samples and those provided by the  
28 classifier (Ng et al. (2001)).

29 Beside class labels, there is another kind of user supervision information  
30 called the pairwise constraints (Bar-Hillel et al. (2005)). It consists to simply  
31 specify whether a pair of data samples must be regrouped together (must-  
32 link constraints) or cannot be regrouped together (cannot-link constraints).  
33 Zhang et al. (Zhang et al. (2008)) propose to evaluate the feature relevance  
34 by scores which only take into account these constraints. Zhao et al. de-  
35 fine another score which uses both the pairwise constraints defined by the  
36 user and the unlabeled nearest neighbors of the samples (Zhao et al. (2008)).  
37 However, this score considers the neighbors of each sample without explicitly

38 taking into account its local density property.

39 These authors have experimentally shown that the features selected thanks  
40 to these constraint-based scores, may provide results which are compara-  
41 ble with those given by supervised approaches. Unfortunately, these scores  
42 strongly depend on the given constraint subsets built by the user. So, when  
43 the user slightly modifies the constraint subset, the feature scores may also  
44 change.

45

46 In this paper, we first present a review of the feature selection scores. We  
47 then propose a semi-supervised score which uses both pairwise constraints  
48 and the local properties of the unlabeled data. We experimentally demon-  
49 strate that this score, thanks to the contribution of the unlabeled data, is  
50 less sensitive to constraint changes than the classical constraint scores, while  
51 providing satisfying classification results.

52 Previous works compare the performances of the feature scores by consider-  
53 ing the accuracy rates obtained by a classifier operating in the selected fea-  
54 ture space. In order to measure the sensitiveness of the scores to constraint  
55 changes, we could estimate the dispersion of accuracy rates with respect to  
56 different subsets of constraints. Though, this evaluation depends on the be-

57 havior of the used classifier. So, we propose to only examine the dispersion  
58 of the feature ranks provided by the examined scores thanks to the Kendall's  
59 coefficient (Grzegorzewski (2006)).

60 The paper is organized as follows. In section 2, we review different supervised  
61 and unsupervised feature selection scores. Then, we introduce the spectral  
62 semi-supervised score in section 3. Recent constraint scores are detailed in  
63 section 4 and our semi-supervised constraint score is presented in section 5.  
64 In section 6, we study the relationships between the constraints given by the  
65 user and the feature ranks obtained by the different scores. Comparative  
66 experimental results are provided in section 7 in order to assess the efficiency  
67 of our semi-supervised constraint score.

## 68 2. Supervised and unsupervised feature selection

69 Given a dataset of  $n$  samples defined in a  $d$ -dimensional feature space,  
70 let us denote  $\mathcal{X} = (x_{ir})$   $i = 1, \dots, n; r = 1, \dots, d$ ; the associated data matrix  
71 where  $x_{ir}$  is the  $r^{\text{th}}$  feature value of the  $i^{\text{th}}$  data. Each of the  $n$  rows of the  
72 matrix  $\mathcal{X}$  represents a data sample  $x_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$ , while each of the  
73  $d$  columns of  $\mathcal{X}$  defines the feature values  $f_r = (x_{1r}, \dots, x_{nr})^T \in \mathbb{R}^n$ .

74 *2.1. Supervised feature selection*

75 The principle of supervised feature selection consists in examining the  
76 correlation between projected data samples and their class labels on each  
77 feature axis. It looks for features on which the classes are compact and far  
78 from each others. For this purpose, one uses the well known Fisher criterion  
79 to evaluate the feature relevance (Bishop (1996)).

80 By considering the sample coordinates on the feature  $f_r$ , each class  $\omega$ ,  $\omega=1$ ,  
81 ...,  $c$ , populated with  $n_\omega$  labeled samples is characterized by its mean  $\mu_{\omega r}$  and  
82 its variance  $\sigma_{\omega r}^2$ . Moreover, let us denote  $\mu_r$  the mean of all data samples on  
83 the feature  $f_r$ .

84 The Fisher score  $F_r$  used to evaluate the relevance of the feature  $f_r$  is defined  
85 by:

$$F_r = \frac{\sum_{\omega=1}^c n_\omega (\mu_{\omega r} - \mu_r)^2}{\sum_{\omega=1}^c n_\omega \sigma_{\omega r}^2}. \quad (1)$$

86 In order to select the most relevant features, they are sorted according to the  
87 decreasing order of their Fisher score  $F_r$ .

88 *2.2. Unsupervised feature selection*

89 Unsupervised feature selection consists in evaluating the relevance of each  
90 feature by examining the dispersion of the data samples projected on its axis.  
91 A feature is considered as being relevant when the data samples projected  
92 on this feature axis are scattered as much as possible.  
93 So, the variance score  $V_r$  is used to evaluate the relevance of the feature  $f_r$ :

$$V_r = \frac{1}{n} \sum_{i=1}^n (x_{ir} - \mu_r)^2. \quad (2)$$

94 The features are sorted according to the decreasing order of  $V_r$ , in order  
95 to select the most relevant ones.

96

97 Rather than measuring the data dispersion along a feature axis, one ex-  
98 amines the local properties of the data. The basic idea is to assume that  
99 the input data pairwise distances are preserved in the relevant feature space.  
100 So, similar samples have to be close when they are projected on a relevant  
101 feature axis.

102 According to the spectral graph theory (von Luxburg. (2007)), data samples  
103  $x_i, i = 1, \dots, n$  are represented by  $n$  nodes in a graph structure. The edge  
104 between two connected nodes  $i$  and  $j$  is weighted by a similarity level defined



105 by:

$$s_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2t^2}\right), \quad (3)$$

106 where  $\|x_i - x_j\|^2$  represents the squared euclidean distance between  $x_i$   
107 and  $x_j$  in the  $d$ -dimensional initial feature space (von Luxburg. (2007)). The  
108 parameter  $t$  has to be tuned in order to represent the local dispersion of the  
109 data (Zelink-Manor and Perona (2005)).

110 The weights of the graph are represented by a similarity matrix  $S$  ( $n \times n$ ).  
111 From  $S$ , we calculate the Laplacian matrix defined as  $L = D - S$ , where  $D$   
112 ( $n \times n$ ) is the diagonal matrix,  $D_{ii} = \sum_{j=1}^n s_{ij}$ . It is interesting to note that  
113 the degree  $D_{ii}$  of a node  $i$  can be considered as a local density measure at  $x_i$ .

114

115 In an unsupervised context, He et al. assume that the projections (coor-  
116 dinates) of similar data on the examined feature axis have to be as close as  
117 possible (He et al. (2005)). They propose to compute the Laplacian score  $L_r$   
118 of a feature  $f_r$  as:

$$L_r = \frac{\sum_{ij} (x_{ir} - x_{jr})^2 s_{ij}}{\sum_i (x_{ir} - \bar{f}_r) D_{ii}}. \quad (4)$$

119 It is easy to demonstrate that:

$$L_r = \frac{\tilde{f}_r^T L \tilde{f}_r}{\tilde{f}_r^T D \tilde{f}_r} \quad (5)$$

120

121 where  $\tilde{f}_r = f_r - \bar{f}_r$  and  $\bar{f}_r = \frac{\sum_{i=1}^n x_{ir} D_{ii}}{\sum_{i=1}^n D_{ii}}$ .

122 By considering  $D_{ii}$  as a density probability measure,  $\bar{f}_r$  is the weighted fea-  
123 ture average.

124 In order to select relevant features, they are sorted according to the ascend-  
125 ing order of  $L_r$ .

126 He et al. have experimentally demonstrated that the classifier operating in  
127 the feature space selected by the Laplacian score  $L_r$ , outperforms the classi-  
128 fier operating in the feature space selected by the variance score  $V_r$ . Indeed,  
129  $L_r$  takes into account the locality structure of the data samples.

### 130 3. Semi-supervised feature selection

131 In many applications, we have huge unlabeled and a few labeled data  
132 samples. Indeed, labeling all the data samples by the user is time consuming  
133 and fastidious. In that context, the labeled data subset is usually too small  
134 to carry sufficient information for the supervised selection while unsupervised

135 approaches ignore this label information which could be yet interesting for  
 136 feature selection.

137 That is the reason why more interest has been addressed for a new challenge  
 138 in feature selection called "small labeled-sample problem". Semi-supervised  
 139 feature selection methods bring slutions by considering both labeled and un-  
 140 labeled data subsets.

141 Zhao et al. propose to couple the Laplacian score with normalized mutual  
 142 information (NMI) in order to introduce a new semi-supervised feature se-  
 143 lection score (Zhao and Liu (2007a); Zhao and Liu (2007c)). This score  
 144 compares the initial labels of the data samples with the labels provided by a  
 145 classifier operating with the examined feature  $f_r$ . It is defined as:

$$M_r = \alpha \frac{\sum_{ij} (x_{ir} - x_{jr})^2 s_{ij}}{\sum_i (x_{ir} - \bar{f}_r) D_{ii}} + (1 - \alpha) \left( 1 - NMI(\hat{f}_r, y) \right). \quad (6)$$

146 where  $\hat{f}_r$  is the cluster indicator of the samples generated by the classifier and  
 147  $y$  is the initial class label vector (Zhao and Liu (2007a)). The first term of  
 148 equation (6) calculates the Laplacian score of the feature  $f_r$ , while the second  
 149 term estimates the corresponding classification error of  $\hat{f}_r$  according to the  
 150 labeled data. The term  $\alpha$  is a regularization parameter set to 0.1 by (Zhao  
 151 and Liu (2007a)), in order to favor the contribution of the labeled data. That

152 is the reason why the selected features by using equation (6) mainly depend  
153 on the labeling decision achieved by the classifier.

#### 154 **4. Constraint scores**

155 The prior knowledge about the data can be represented according to two  
156 different ways: class labels and pairwise constraints. Class labels require  
157 to have detailed information about the classes and to precisely indicate the  
158 label of each data sample. Pairwise constraints simply mention for some  
159 pairs of data samples that they are similar, i.e. must be regrouped together  
160 (must-link constraints), or that they are dissimilar, i.e. cannot be regrouped  
161 together (cannot-link constraints).

162 The user has to build the subset  $\mathcal{M}$  of must-link constraints and the subset  
163  $\mathcal{C}$  of cannot-link constraints defined as:

$$164 \mathcal{M} = \{(x_i, x_j), \text{ such as } x_i \text{ and } x_j \text{ must be linked}\}.$$

$$165 \mathcal{C} = \{(x_i, x_j), \text{ such as } x_i \text{ and } x_j \text{ cannot be linked}\}.$$

166 The cardinals of these subsets are usually much lower than the number  $\binom{n}{2}$   
167 of all possible pairwise constraints.

168 These pairwise constraints are easier to be obtained by the user than the  
169 class labels. They simply formalize that two data samples belong or not to

170 the same class without detailed information about the classes in presence.  
171 Indeed, labeled data samples can be transformed into must-link and cannot-  
172 link constraints but not vice versa. It consists in putting must-link constraint  
173 between two data samples which share the same label and cannot-link con-  
174 straint between two data samples sharing different labels.

175 Zhang et al. have recently proposed a constraint feature selection scheme  
176 which uses only a subset of must-link and cannot-link constraints (Zhang  
177 et al. (2008)).

178 In the context of the spectral theory (von Luxburg. (2007)), two specific  
179 graphs are built:

- 180 • The must-link graph  $G^{\mathcal{M}}$  where a connection is established between  
181 two nodes  $i$  and  $j$  if there is a must-link constraint between their cor-  
182 responding samples (nodes)  $x_i$  and  $x_j$ .
- 183 • The cannot-link graph  $G^{\mathcal{C}}$  where a connection is established between  
184 two nodes  $i$  and  $j$  if there is a cannot-link constraint between their  
185 corresponding samples (nodes)  $x_i$  and  $x_j$ .

186 The connection weights between two nodes of the graphs  $G^{\mathcal{M}}$  and  $G^{\mathcal{C}}$  are  
187 respectively stored by the similarity matrices  $S^{\mathcal{M}}$  ( $n \times n$ ) and  $S^{\mathcal{C}}$  ( $n \times n$ ),

188 and are given by:

189  $s_{ij}^{\mathcal{M}} =$

$$\begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{M} \text{ or } (x_j, x_i) \in \mathcal{M} \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

190

191  $s_{ij}^{\mathcal{C}} =$

$$\begin{cases} 1 & \text{if } (x_i, x_j) \in \mathcal{C} \text{ or } (x_j, x_i) \in \mathcal{C} \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

192 The matrices  $S^{\mathcal{M}}$  and  $S^{\mathcal{C}}$  are used to define the constraint Laplacian  
 193 matrices  $L^{\mathcal{M}} = D^{\mathcal{M}} - S^{\mathcal{M}}$  and  $L^{\mathcal{C}} = D^{\mathcal{C}} - S^{\mathcal{C}}$ , where  $D^{\mathcal{M}}$  and  $D^{\mathcal{C}}$  are the  
 194 degree matrices defined by  $D_{ii}^{\mathcal{M}} = \sum_{j=1}^n s_{ij}^{\mathcal{M}}$  and  $D_{ii}^{\mathcal{C}} = \sum_{j=1}^n s_{ij}^{\mathcal{C}}$ .

195 In order to measure the constraint preserving ability of the feature  $f_r$ , Zhang  
 196 et al. define two constraint scores  $C_r^1$  and  $C_r^2$ :

$$C_r^1 = \frac{\sum_{ij} (x_{ir} - x_{jr})^2 s_{ij}^{\mathcal{M}}}{\sum_{ij} (x_{ir} - x_{jr})^2 s_{ij}^{\mathcal{C}}} = \frac{f_r^T L^{\mathcal{M}} f_r}{f_r^T L^{\mathcal{C}} f_r}, \quad (9)$$

197

198

$$\begin{aligned}
C_r^2 &= \sum_{ij} (x_{ir} - x_{jr})^2 s_{ij}^{\mathcal{M}} - \lambda \sum_{ij} (x_{ir} - x_{jr})^2 s_{ij}^{\mathcal{C}} \\
&= f_r^T L^{\mathcal{M}} f_r - \lambda f_r^T L^{\mathcal{C}} f_r,
\end{aligned} \tag{10}$$

199 where  $\lambda$  is a regularization parameter used to balance the contribution of  
200 the two constraints terms of  $C_r^2$ . Must-link constraints are favored by setting  
201  $0 < \lambda < 1$ .

202 The lower these two scores are, the more relevant the feature is. Zhang et al.  
203 have experimentally shown that the features selected by  $C^1$  and  $C^2$  provide  
204 similar performances when  $\lambda$  is well balanced.

## 205 5. Constraint scores for semi-supervised feature selection

206 The scores presented in section 4 use only the available constraints and  
207 do not take into account the unlabeled data contribution. Zhao et al. define  
208 another score which uses both unlabeled data and pairwise constraints in  
209 order to retrieve both locality properties and discriminating structures in  
210 the data samples (Zhao et al. (2008)). They build a new graph  $G^{\mathcal{W}}$  which  
211 connects samples having high probability of sharing the same label:

- 212 •  $G^{\mathcal{W}}$  is the within-class graph: two nodes  $i$  and  $j$  are connected if  $(x_i, x_j)$   
213 or  $(x_j, x_i)$  belongs to  $\mathcal{M}$ , or if the two samples are unlabeled but they

214 are sufficiently close to each other (by using the  $k$ -nearest neighbor  
 215 graph)

216 The edges in the graphs  $G^{\mathcal{W}}$  are weighted by using the similarity matrix  $S^{\mathcal{W}}$   
 217 ( $n \times n$ ) and are expressed as:

218  $s_{ij}^{\mathcal{W}} =$

219

$$\left\{ \begin{array}{l} \gamma \text{ if } (x_i, x_j) \in \mathcal{M} \text{ or } (x_j, x_i) \in \mathcal{M} \\ 1 \text{ if } x_i \text{ or } x_j \text{ is unlabeled} \\ \text{but node } i \in KNN(j) \text{ or node } j \in KNN(i) \\ 0 \text{ otherwise} \end{array} \right. \quad (11)$$

220 where  $\gamma$  is a constant parameter which has been empirically set to 100 in  
 221 (Zhao et al. (2008)).

222 Zhao et al. introduce a Laplacian score, called the locality sensitive discrim-  
 223 inant analysis score and defined as:

$$C_r^3 = \frac{\sum_{i,j} (x_{ir} - x_{jr})^2 s_{ij}^{\mathcal{W}}}{\sum_{i,j} (x_{ir} - x_{jr})^2 s_{ij}^{\mathcal{C}}} = \frac{f_r^T L^{\mathcal{W}} f_r}{f_r^T L^{\mathcal{C}} f_r}, \quad (12)$$



224 where  $L^{\mathcal{W}} = D^{\mathcal{W}} - S^{\mathcal{W}}$ ,  $D^{\mathcal{W}}$  being the degree matrix defined by  $D_{ii}^{\mathcal{W}} =$   
 225  $\sum_{j=1}^n s_{ij}^{\mathcal{W}}$ . This score implicitly takes into account the unlabeled data but  
 226 favors pairs of must-link data by assigning them high weights in the matrix  
 227  $S^{\mathcal{W}}$ . Moreover, the similarity matrix  $S^{\mathcal{W}}$  represents the links between the  
 228  $k$ -nearest neighbors of the data by binary weighting them. By mainly con-  
 229 sidering the must-link constraints,  $C^3$  is very close to  $C^2$ , and both neglect  
 230 the unlabeled data samples.

231 Though, taking into account the unlabeled data samples should catch the  
 232 data structure and make less sensitive a feature score against the given con-  
 233 straint subset. That leads us to propose another semi-supervised constraint  
 234 score which is less sensitive to the constraints chosen by the user. Given the  
 235 matrices  $S$ ,  $S^{\mathcal{M}}$  and  $S^{\mathcal{C}}$ , the semi-supervised constraint score is defined as :

$$C_r^4 = \frac{\tilde{f}_r^T L \tilde{f}_r}{\tilde{f}_r^T D \tilde{f}_r} \cdot \frac{f_r^T L^{\mathcal{M}} f_r}{f_r^T L^{\mathcal{C}} f_r}. \quad (13)$$

236 The proposed score  $C_r^4$  is the simple product between the Laplacien score  
 237  $L_r$  (see equation (4)) processed with the unlabeled data and the constraint  
 238 score  $C_r^1$  (see equation (9)) defined by Zhang:

$$C_r^4 = L_r \cdot C_r^1, \quad (14)$$

239 So, it takes into account both the unlabeled data thanks to  $L_r$  and the  
240 available constraints thanks to  $C_r^1$  in order to evaluate the relevance of the  
241 feature  $f_r$ . As for the other scores, the features are ranked in ascending order  
242 according to score  $C^4$  in order to select the most relevant ones.

## 243 6. Constraint subset influence in feature selection process

244 The constraint scores introduced above evaluate the relevance of the fea-  
245 tures based on the must-link and cannot-link constraints. However, these  
246 scores strongly depend on the given constraint subsets  $\mathcal{M}$  and  $\mathcal{C}$ . Indeed,  
247 changing the subset of available constraints could lead to a large change in  
248 the feature ranks and so, in the selected features. To illustrate this problem,  
249 we examine the following toy example.

### 250 6.1. Toy example

251 Let us consider four 3-dimensional data samples:

252  $A(-3, -1, 1)$ ;  $B(-3, 1, 1)$ ;  $C(-1, -1, 1)$  and  $D(1, -3, -1)$ .

253  $A$ ,  $B$  and  $C$  belong to the first class whereas  $D$  is assigned to the second  
254 class. We can see that the feature  $f_3$  is the single feature which corresponds  
255 to the class label (1 for the first class,  $-1$  for the second class). So, an efficient

256 feature selection algorithm should identify the feature  $f_3$  as the most relevant  
257 one.

258 Let us consider that the user builds one single must-link constraint and one  
259 single cannot-link constraint from the labeled samples. The constraint scores  
260 introduced in sections 4 and 5 are used to rank the 3 features based on these  
261 constraints.

262 The different feature ranks are shown in table 1. Pairs  $\{(A,B)\}$ ,  $\{(A,C)\}$  and  
263  $\{(B,C)\}$  are the single must-link constraints that can be built from the data  
264 samples. Pairs  $\{(A,D)\}$ ,  $\{(B,D)\}$  and  $\{(C,D)\}$  are the single cannot-link  
265 constraints that can be also built from the data samples. So, there are 9  
266 constraint combinations which correspond to 9 cells in table 1. The first, the  
267 second, the third and the fourth row of each cell represent the ranks of the 3  
268 features given by  $C^1$ ,  $C^2$  ( $\alpha$  is fixed to 0.1 as suggested by the authors),  $C^3$   
269 and  $C^4$  scores, respectively. The sign '=' between two features means that  
270 their scores are equal.

271 For example, let us examine the cell which corresponds to the must-link  
272 constraint  $\{(A,C)\}$  and the cannot-link constraint  $\{(B,D)\}$ . Scores  $C^1$  and  
273  $C^4$  select the features  $f_2$  and  $f_3$  as the best features. and the feature  $f_1$  as the  
274 third feature. Scores  $C^2$  and  $C^3$  sort feature  $f_2$  as the best feature, feature

$f_3$  as the second feature and feature  $f_1$  as the third feature.

$\mathcal{C}/\mathcal{M}$		$\{(A,B)\}$	$\{(A,C)\}$	$\{(B,C)\}$
$\{(A,D)\}$	$C^1$	$\mathbf{f_3 = f_1, f_2}$	$\mathbf{f_3 = f_2, f_1}$	$\mathbf{f_3, f_1, f_2}$
	$C^2$	$\mathbf{f_1, f_3, f_2}$	$f_3 = f_2, f_1$	$f_3, f_1, f_2$
	$C^3$	$f_1, f_3, f_2$	$\mathbf{f_3, f_2, f_1}$	$f_3, f_1, f_2$
	$C^4$	$f_3 = f_1, f_2$	$f_3 = f_2, f_1$	$f_3, f_1, f_2$
$\{(B,D)\}$	$C^1$	$f_3 = f_1, f_2$	$f_3 = f_2, f_1$	$\mathbf{f_3, f_1 = f_2}$
	$C^2$	$f_1, f_3, f_2$	$\mathbf{f_2, f_3, f_1}$	$f_3, f_1 = f_2$
	$C^3$	$f_1, f_3, f_2$	$f_2, f_3, f_1$	$f_3, f_2, f_1$
	$C^4$	$f_3 = f_1, f_2$	$f_3 = f_2, f_1$	$f_3, f_1, f_2$
$\{(C,D)\}$	$C^1$	$f_3 = f_1, f_2$	$f_3 = f_2, f_1$	$f_3, f_1 = f_2$
	$C^2$	$f_3 = f_1, f_2$	$f_3 = f_2, f_1$	$f_3, f_1 = f_2$
	$C^3$	$f_3, f_1, f_2$	$f_3, f_2, f_1$	$f_3, f_2, f_1$
	$C^4$	$f_3 = f_1, f_2$	$f_3 = f_2, f_1$	$f_3, f_1, f_2$

Table 1: Feature ranks by semi-supervised scores using the chosen constraints.

275

276 First, table 1 shows that the feature  $f_3$  is always ranked as the first

277 feature by  $C^1$  and  $C^4$  whatever the given constraint subset.  $C^2$  and  $C^3$  do

278 not retain feature  $f_3$  as the first feature three times. This table also shows

279 that 7 different ranks (indicated in bold in table 1) of the 3 features are  
280 obtained by comparing each of the 4 examined scores for the 9 constraint  
281 combinations. This simple example clearly shows that the features ranks  
282 deduced from these scores strongly depend on the chosen constraints. So, we  
283 propose to measure how the feature ranks vary with respect to the subset of  
284 constraints.

## 285 *6.2. Kendall's coefficient*

286 The performances of the examined scores are generally compared by mea-  
287 suring the accuracy rates obtained from well classified data projected on the  
288 selected features. To measure the dependence of the scores on the given  
289 constraint subsets, we could estimate the disparity of the accuracy rates ob-  
290 tained with different constraint subsets. This evaluation requires to define a  
291 labeling decision step, such as  $k$ -nearest neighbor classifier, which operates in  
292 the selected feature space. Since this decision step influences the quality of  
293 classification, the comparison between the score performances may depend  
294 on the used classifier.

295 We prefer to examine only the feature ranks deduced from the scores, so that  
296 our study is not corrupted by the decision step. More precisely, we study the

297 concordance between feature ranks when the constraint subset changes.

298

299 Given the sample set  $\mathcal{X} = (x_1, \dots, x_n)$  and the associated class label of each  
300 data sample, we randomly pick up a subset  $\mathcal{S}_q$  of pairwise constraints ( $\mathcal{S}_q =$   
301  $\mathcal{M}_q \cup \mathcal{C}_q$ ). Then, we rank the  $d$  features  $f_r$ , according to the different con-  
302 straint scores  $C_r^*$  detailed in sections 4 and 5 ( $* = 1, 2, 3, 4$ ).

303 Let us denote  $R_{qr}^*$  the rank of the feature  $f_r$  when the score  $C_r^*$  considers the  
304 constraint subset  $\mathcal{S}_q$ . In order to evaluate the influence of constraint subset,  
305 the feature selection is run  $p$  times. For a score  $C^*$ , the ranks of the  $d$  features  
306 obtained with the  $p$  different subsets  $\mathcal{S}_q$  are represented by the matrix  $R^*$   
307 ( $p \times d$ ):

308

$$309 \quad R^* = \begin{bmatrix} R_{11}^* & R_{12}^* & \dots & R_{1d}^* \\ R_{21}^* & R_{22}^* & \dots & R_{2d}^* \\ \dots & \dots & \dots & \dots \\ R_{p1}^* & R_{p2}^* & \dots & R_{pd}^* \end{bmatrix}$$

310

311 The  $q^{th}$  row of  $R^*$  represents the  $d$  feature ranks by using the subset  $\mathcal{S}_q$   
312 ( $q = 1, \dots, p$ ) while the  $r^{th}$  column represents the ranks of the feature  $f_r$  by  
313 using the  $p$  different constraint subsets. Therefore, each row of  $R^*$  is a per-

314 mutation of the  $d$  feature ranks which depends on the available constraint  
 315 subset  $S_q$ .

316 We use the Kendall's coefficient to measure the concordance or agreement  
 317 between the feature ranks with  $p$  constraint subsets (Grzegorzewski (2006)).

318 The Kendall's coefficient  $K^*$  takes into account the different rows of the  
 319 matrix  $R^*$  and is defined as (Siegel and Castellan (1988)):

$$K^* = \frac{12\Delta^*}{p^2(d^3 - d) - p\tau^*}, \quad (15)$$

320 where  $\Delta^* = \sum_{r=1}^d (R_r^* - \bar{R}^*)^2$ ,  $R_r^* = \sum_{q=1}^p R_{qr}^*$ ,  $\bar{R}^* = \frac{1}{d} \sum_{r=1}^d R_r^*$  and  $\tau^* =$   
 321  $\sum_{v=1}^m (\tau_v^{*3} - \tau_v^*)$

322 The term  $\tau_v^*$  is the number of tied ranks in each of the  $m$  groups of ties in  
 323  $R^*$ . The sum  $\tau^*$  is computed over all the groups of ties found in all  $p$  rows  
 324 of the table  $R^*$ .

325 The Kendall's coefficient which measures the dispersion  $\Delta^*$  of the feature  
 326 ranks, ranges from 0 (no agreement) to 1 (complete agreement).

### 327 6.3. Kendall's coefficient for the toy example

328 Let us examine the toy example of section 6.1. We propose to compute  
 329 the Kendall's coefficient from table 1 in order to measure the concordance of

330 the feature ranks obtained by each examined constraint score. The Kendall's  
331 coefficients  $K^1$  (see eq. (9)),  $K^2$  (see eq. (10)),  $K^3$  (see eq. (12)) and  $K^4$   
332 (see eq. (13)) are respectively 0.4325, 0.2258, 0.3333 and 0.4333.

333 These low values reflect the dependence of the feature ranks on the available  
334 constraints subsets.

335 This example shows that the Kendall's coefficient  $K^4$  of  $C^4$  is slightly higher  
336 than the other ones. The objective of this simple example, with low sample  
337 population, is not to compare the Kendall's coefficient of the different scores,  
338 but to arise that the selected features depend on the given constraint subset.  
339 This conclusion concurs with that of Sun et al. (Suna and Zhang (2010)).

## 340 7. Comparative experimental results

341 In this section, we first measure the sensitivity of the scores against the  
342 given constraints. So, we compare the Kendall's coefficient obtained by the  
343 tested scores ( $C^1$ ,  $C^2$ ,  $C^3$ ,  $C^4$ ).

344 We also compare the performances obtained by a classifier operating in the  
345 feature space selected by  $C^4$  score and those obtained with the classical  
346 feature selection methods: Fisher score (supervised), Laplacian score (un-  
347 supervised),  $C^1$ ,  $C^2$  (constraint scores) and  $C^3$  (semi-supervised constraint



348 score).

### 349 *7.1. Examined databases*

350 For this purpose, our experiments are achieved with six well known and  
351 largely used benchmark databases, and more precisely the 'Wine', 'Image  
352 segmentation' and 'Vehicle' databases from the UCI repository (Blake et al.  
353 (1998)), the face database 'ORL' (Samaria and Hartert (1994)) and the  
354 two gene expression databases, i.e., 'Colon Cancer'(Alon et al. (1999)) and  
355 'Leukemia'(Golub et al. (1999)). These databases have been retained since  
356 the features are numeric and since the label information of each data sample  
357 is clearly defined.

358 In our experiments, we first normalize the features between 0 and 1, so that  
359 different features have the same scale. For each dataset, we follow an Holdout  
360 partition and choose the first half of samples from each class as the training  
361 data and the remaining data for testing.

### 362 *7.2. Results on UCI datasets and ORL database*

363 First, we achieve experiments on 3 UCI datasets and the the 'ORL' face  
364 database. Here is a brief description of the four considered databases:

365 • **'Wine' database**

366 This database contains 178 13-dimensional data ( $d=13$ ) regrouped into  
367 3 classes having 59, 71 and 48 instances, respectively. We randomly  
368 select 30, 36 and 24 data samples from each class to build the training  
369 data. The remaining data are organized as the test data subset.

370 • **'Image segmentation' database**

371 This database contains 210 19-dimensional data ( $d=19$ ) regrouped into  
372 7 classes, each class having 30 instances. We randomly select 15 data  
373 samples from each class to build the training data and the remaining  
374 data constitute the test data subset.

375 • **'Vehicle' database**

376 This database contains 846 18-dimensional data ( $d=18$ ) regrouped into  
377 4 classes, having 212, 217, 218 and 199 instances, respectively. We  
378 randomly select 106, 109, 109 and 100 data samples from each class to  
379 build the training data.

380 • **'ORL' database**

381 The 'ORL' database (Olivetti Research Laboratory) contains a set of  
382 face images representing 40 distinct subjects. There are 10 different



Figure 1: Sample face images from the ORL database (2 subjects).

383 images per subject, so that the database contains 400 images. For each  
384 subject, the images have been acquired according to different condi-  
385 tions: lighting, facial expressions (open / closed eyes, smiling / not  
386 smiling) and facial details (glasses / no glasses) (see figure 1).

387 In our experiments, original images have been normalized (in scale and  
388 orientation) so that the two eyes are aligned at the same horizontal  
389 position. Then, the facial areas have been cropped in order to build  
390 images of size  $32 \times 32$  pixels. Thus, each image can be represented by  
391 a 1024-dimensional sample data. The gray level of each pixel is quan-  
392 tified with 256 levels.

393 We randomly select 5 images from each class (subject) to build the  
394 training data. The remaining data are organized as the test data sub-  
395 set.

396 7.2.1. Kendall's coefficient results

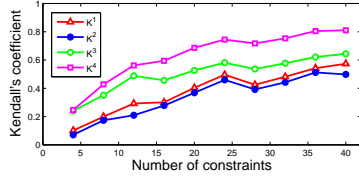
397 In our experiments, the feature selection is performed on the training data  
398 and features are ranked according to the different scores. At each feature se-  
399 lection run  $q$ ,  $q=1, \dots, p$ , we simulate the generation of pairwise constraints  
400 as follow: we randomly select pairs of samples from the training data and  
401 create must-link or cannot-link constraints depending on whether the under-  
402 lying classes of the two samples are the same or different. We iterate this  
403 scheme until we obtain  $\frac{\text{card}(S_q)}{2}$  must-link constraints and  $\frac{\text{card}(S_q)}{2}$  cannot-link  
404 constraints.

405 This operation is repeated over  $p = 100$  runs in order to measure the  
406 Kendall's coefficients  $K^1$ ,  $K^2$ ,  $K^3$  and  $K^4$ .

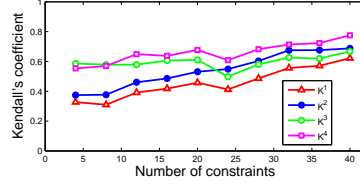
407 We achieve experiments for different cardinals of  $S_q$  ranging from 4 (2 must-  
408 link and 2 cannot-link) to 40 constraints as Zhang et al. do.

409 Figure 2 shows the Kendall's coefficients  $K^1$ ,  $K^2$ ,  $K^3$  and  $K^4$  calculated  
410 over  $p = 100$  runs for different numbers of constraints. At each run, the  
411 same given constraint subset is of course considered by the four tested cri-  
412 teria. The low values of  $K^1$ ,  $K^2$ ,  $K^3$  and  $K^4$  show clearly that the selected  
413 features using  $C^1$ ,  $C^2$ ,  $C^3$  and  $C^4$  depend on the constraint subsets.

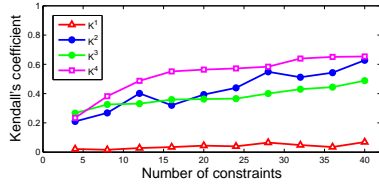
414 By examining the different curves of figure 2, we see that the different coeffi-



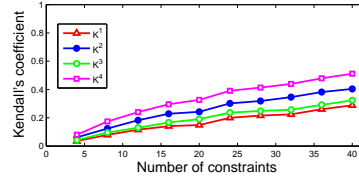
(a) 'Wine'.



(b) 'Image segmentation'.



(c) 'Vehicle'.



(d) 'ORL'.

Figure 2: Kendall's coefficient for different number  $S_q$  of constraints on the four databases.

415 cients  $K^1$ ,  $K^2$ ,  $K^3$  and  $K^4$  increase with the number  $card(S_q)$  of constraints.

416 The higher the number  $card(S_q)$  of constraints is, the more complete the su-

417 pervision information is. Ideed, when  $card(S_q)$  is high, the semi-supervised

418 learning context tends to become a supervised learning context. That ex-

419 plains why the agreement of feature ranks increases when  $card(S_q)$  increases

420 in the different curves of figure 2.

421 We can also notice that  $K^4$  has the highest values for the different cardinals

422 of constraint subsets in the four databases. This shows that our score  $C^4$  is

423 less sensitive to the must-link and cannot-link constraints built by the user

424 than the classical scores.

425 We also observe that the values of the Kendall's coefficients on the 'ORL'  
426 database are lower than those obtained with the other databases. Indeed,  
427 the number of features extracted from the 'ORL' database (1024 features) is  
428 bigger than that of the other databases.

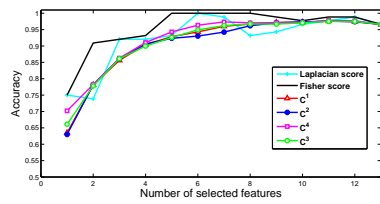
429 Zhao and Zhang have experimentally shown that by picking-up a small num-  
430 ber  $card(\mathcal{S}_q)$  of constraints, their scores allow to select the features which  
431 carry a good discriminating power. Figure 2 shows that, between two differ-  
432 ent runs, the feature ranks change, and so the selected features change when  
433  $card(\mathcal{S}_q)$  is low. So, the quality of data discrimination would strongly vary  
434 with respect to the chosen constraint subsets.

### 435 *7.2.2. Comparison of the performances*

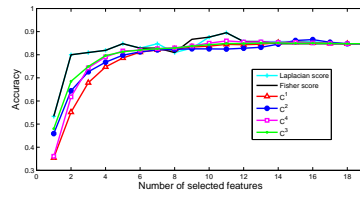
436 We also compare the performances obtained by the nearest neighbor clas-  
437 sifier which operates in the feature space selected thanks to the considered  
438 supervised, unsupervised and semi-supervised feature scores.

439 The classification accuracies of the test data are used to evaluate the perfor-  
440 mance of each criterion.

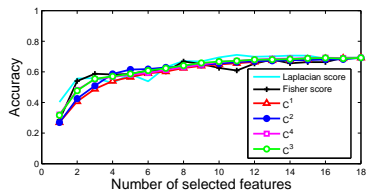
441 As for measuring the Kendall's coefficients, the rates of good classification



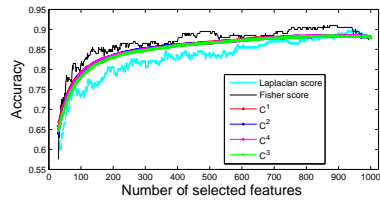
(a) 'Wine'.



(b) 'Image segmentation'.



(c) 'Vehicle'.



(d) 'ORL'.

Figure 3: Accuracy vs. different number of selected features on the four datasets. 10 pairwise constraints including 5 must-link and 5 cannot link are used.

442 are averaged over  $p = 100$  runs with different generations of constraints.

443 Figure 3 shows the plots of accuracy vs. the desired number of selected fea-

444 tures on the 'Wine' 3(a), 'Image segmentation' 3(b), 'Vehicle' 3(c) and 'ORL'

445 3(d) databases, respectively when  $card(\mathcal{S}_q)$  is set to 10 .

446 From figure 3 , we can see that the classification rates obtained with con-

447 straint scores ( $C^1$ ,  $C^2$ ,  $C^3$  and  $C^4$ ) range between those obtained with the un-

448 supervised method (Laplacian score) and those obtained with the supervised

449 method (Fisher score). Since  $card(\mathcal{S}_q)$  is set to 10, these semi-supervised

450 scores use a bit more supervision information than Laplacian score, but much

451 fewer than Fisher Score.

452 So, these results confirm that a classification scheme reaches higher per-

453 formance thanks to a supervised learning than thanks to a semi-supervised

454 learning. They also confirm that taking into account a few constraints allows

455 to improve classification results, compared with the unsupervised learning.

456 However, the curves of  $C^1$ ,  $C^2$ ,  $C^3$  and  $C^4$  are confused in figure 3. It is also

457 difficult to compare the performances of these scores since they are averaged

458 over 100 runs with different generations of constraints.

459 That leads us to compare these scores while examining their accuracies

460 at each of the 100 runs. For a fixed number of selected features, in each of



$card(S_q) / T$	$T^1$	$T^2$	$T^3$	$T^4$
4 constraints	198	238	219	<b>180</b>
10 constraints	195	210	185	<b>162</b>
40 constraints	180	299	168	<b>136</b>

Table 2: The rank sums of the different semi-supervised criteria for different number  $S_q$  of constraints on the 'wine' database.

$card(S_q) / T$	$T^1$	$T^2$	$T^3$	$T^4$
4 constraints	208	228	<b>168</b>	179
10 constraints	228	210	<b>154</b>	183
40 constraints	184	177	153	<b>146</b>

Table 3: The rank sums of the different semi-supervised criteria for different number  $S_q$  of constraints on the 'image' database.

461 the 100 runs, we propose to rank the 4 criteria in descending order of their  
462 accuracy. Let us denote  $rank_q^*$  the rank of the criterion  $C^*$  at the run  $q$ .  
463 This rank takes the values 1, 2, 3 or 4. At each run  $q$ , the method having  
464 the highest accuracy is ranked as 1 and the method with the lowest accuracy  
465 value, is ranked as 4. Methods with the same accuracy have the same rank.  
466 We calculate a rank sum  $T^*$  for each semi-supervised constraint score as  
467 follow:

$card(S_q) / T$	$T^1$	$T^2$	$T^3$	$T^4$
4 constraints	237	239	232	<b>219</b>
10 constraints	271	246	<b>189</b>	240
40 constraints	290	240	208	<b>207</b>

Table 4: The rank sums of the different semi-supervised criteria for different number  $S_q$  of constraints on the 'vehicle' database.

$card(S_q) / T$	$T^1$	$T^2$	$T^3$	$T^4$
4 constraints	<b>194</b>	212	311	200
10 constraints	<b>185</b>	189	250	196
40 constraints	<b>190</b>	202	300	203

Table 5: The rank sums of the different semi-supervised criteria for different number  $S_q$  of constraints on the 'ORL' database.

$$T^* = \sum_{q=1}^{100} rank_q^*, \quad (16)$$

468 where \* is 1, 2, 3 or 4 corresponding to the score  $C^1$ ,  $C^2$ ,  $C^3$  or  $C^4$  respectively.

469 The method with the lowest rank sum is considered as being the score which

470 provides the best results.

471 For the 'Wine', 'Image segmentation', 'Vehicle' and 'ORL' databases, the

472 accuracy of each of the 4 semi-supervised criteria seems to be stable when

473 the number of desired features is higher than 6, 5, 8 and 300, respectively  
474 (see figure 3). So, we propose to calculate the rank total of each of the  
475 semi-supervised criteria by considering the 6 first selected features for the  
476 'Wine database', the 5 first selected features for the 'Image segmentation'  
477 database, the 8 first selected features for the 'Vehicle' database and the 300  
478 first selected ones for the 'ORL' database .

479 Tables 2, 3, 4 and 5 show the rank sum  $T^*$  for different numbers  $S_q$  of  
480 constraints (4, 10 and 40). Each cell indicates the rank sum of the tested  
481 score with the considered number of constraints. From these tables, we can  
482 clearly see that  $T^1$ ,  $T^2$ ,  $T^3$  and  $T^4$  are very close. The features selected  
483 thanks to  $C^4$  provide accuracy rates comparable with those obtained by the  
484 features selected by  $C^1$ ,  $C^2$  and  $C^3$ . Indeed, our score provides the lowest  
485 rank sum  $T$  (indicated in bold) for 6 times over the 12 rows of tables 2, 3, 4  
486 and 5.

### 487 *7.3. Results on gene expression databases*

488 In this subsection, several experiments are carried out on two gene expres-  
489 sion databases, i.e., 'Colon Cancer'(Alon et al. (1999)) and 'Leukemia'(Golub  
490 et al. (1999)).

491     • **'Colon Cancer'**

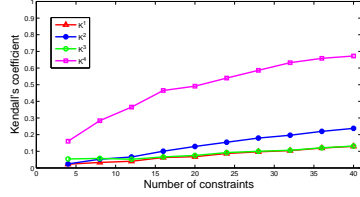
492         This database contains the expression of 2000 genes measured on 62  
493         tissues (40 tumors and 22 normals). We randomly select 20 and 11 data  
494         samples from each class to build the training data. The remaining data  
495         are organized as the test data subset.

496     • **'Leukemia'**

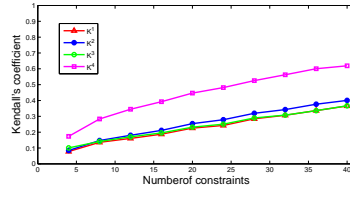
497         This database contains information on gene-expression in samples from  
498         human acute myeloid (AML) and acute lymphoblastic leukemias (ALL).  
499         From the originally measured 6817 probe sets we removed genes that  
500         were not present in at least one sample so a total of 5147 genes are  
501         used in the experiments.

502         Because Leukemia has a predefined partition of the objects into train-  
503         ing (27 ALL and 11 AML) and testing (20 ALL and 14 AML) subsets,  
504         the ensembles on this dataset are performed on the predefined training  
505         and testing sets.

506     Figure 4 shows the Kendall's coefficients  $K^1$ ,  $K^2$ ,  $K^3$  and  $K^4$  calculated over  
507      $p = 100$  runs for different numbers of constraints on the 'Colon Cancer' and  
508     the 'Leukemia' databases. As for the previous databases,  $K^4$  has the high-



(a) 'Colon Cancer'.

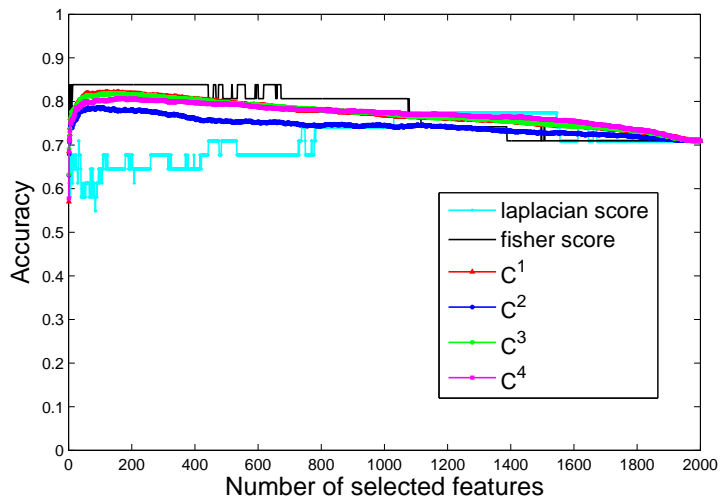


(b) 'Leukemia'.

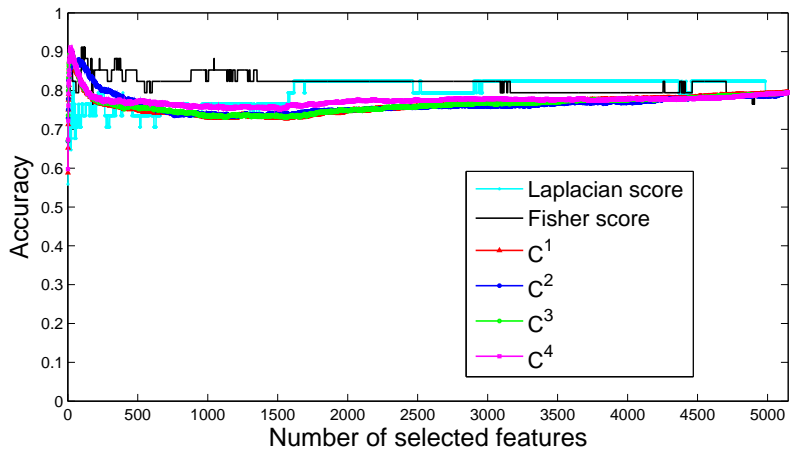
Figure 4: Kendall's coefficient for different number  $S_q$  of constraints on the 'Colon Cancer' and the 'Leukemia' databases.

509 est values for the different cardinals of constraint subsets in these two gene  
 510 expression databases. This confirm the fact that our score  $C^4$  is less sensi-  
 511 tive to the must-link and cannot-link constraints built by the user than the  
 512 classical scores. Moreover, since the number of features is very high, 2000  
 513 for the Colon Cancer database and 5147 for the leukemia one, we can notice  
 514 that the gap between  $K^4$  on a hand and  $K^1$ ,  $K^2$  and  $K^3$  on the other hand  
 515 is higher on these databases than the previous ones (a voir).

516 Figure 5 shows the plots for accuracy vs. different numbers of selected fea-  
 517 tures on Colon Cancer and Leukemia databases.  $card(S_q)$  here is set to 60  
 518 as Sun et al. do (Suna and Zhang (2010)), so 60 pairwise constraints in-  
 519 cluding 30 must-link and 30 cannot-link are used. From figure 5, we can  
 520 see that the classification rates obtained with constraint scores (C1, C2,



(a) 'Colon Cancer'.



(b) 'Leukemia'.

Figure 5: Accuracy vs. different number of selected features on the gene expression-databases. 60 pairwise constraints including 30 must-link and 30 cannot link are used.

521 C3 and C4) range between those obtained with the unsupervised method  
 522 (Laplacian score) and those obtained with the supervised method (Fisher  
 523 score). These semi-supervised constraint scores even outperforms the super-  
 524 vised Fisher score on the 'Colon Cancer' database for a number of selected  
 features more than 1050.

$card(S_q) / T$	$T^1$	$T^2$	$T^3$	$T^4$
4 constraints	<b>148</b>	195	235	158
10 constraints	156	242	203	<b>153</b>
40 constraints	169	286	153	<b>138</b>
60 constraints	157	271	149	<b>144</b>

Table 6: The rank sums of the different semi-supervised criteria for different number  $S_q$  of constraints on the 'Colon Cancer' database.

525

526 Tables 6 and 7 show the rank sum  $T^*$  for different numbers  $S_q$  of constraints  
 527 (4, 10, 40 and 60) on the 'Colon Cancer' and the 'Leukemia' databases re-  
 528 spectively. The rank total of each of the semi-supervised criteria is calculated  
 529 considering the half of the original features of each of the gene expression  
 530 databases (see(Suna and Zhang (2010))).

531 From these tables, we can see that, for the 'Colon Cancer' database, our  
 532 score provides the lowest rank sum T (indicated in bold) for 3 times over the

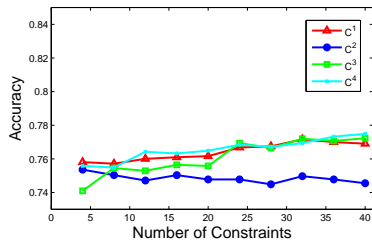
$card(S_q) / T$	$T^1$	$T^2$	$T^3$	$T^4$
4 constraints	184	240	194	<b>182</b>
10 constraints	176	276	150	<b>134</b>
40 constraints	170	196	148	<b>118</b>
60 constraints	186	198	168	<b>126</b>

Table 7: The rank sums of the different semi-supervised criteria for different number  $S_q$  of constraints on the 'Leukemia' database.

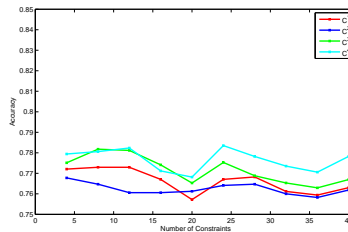
533 4 rows of table 6. For the 'Leukemia database', our score provides the lowest  
534 rank sum  $T$  (indicated in bold) for the different numbers of constraints (4,  
535 10, 40 and 60). These results improve the fact that the features selected  
536 thanks to  $C_4$  provide accuracy rates comparable with those obtained by the  
537 features selected by  $C_1$ ,  $C_2$  and  $C_3$ .

538 Furthermore, figure 6 shows the plot for accuracy under fixed number of se-  
539 lected features ( half of the number of original features) vs. different numbers  
540 of pairwise constraints on the gene expression databases. For almost all the  
541 number of constraints, our score  $C_4$  achieves higher accuracy than  $C_1$ ,  $C_2$   
542 and  $C_3$ .





(a) 'Colon Cancer'.



(b) 'Leukemia'.

Figure 6: Accuracy vs. different number of pairwise constraints (for  $C_1$ ,  $C_2$ ,  $C_3$  and  $C_4$ ) on the gene expression databases. The desired number of selected features is half of the number of original features

## 543 8. Conclusion

544 Constraint scores which use pairwise constraints for semi-supervised fea-  
 545 ture selection have shown good performances of classification. Unfortunately,  
 546 these scores depend on the subset of constraints built by the user since they  
 547 do not take into account the information provided by the unlabeled data. In  
 548 this paper, we propose a new semi-supervised constraint score that considers  
 549 both the pairwise constraints and the local properties of the unlabeled data.  
 550 Moreover, we study the relationships between the features selected by the  
 551 constraint scores and the constraints chosen by the user. We measure the  
 552 sensitiveness of the scores to constraint changes by the Kendall's coefficient.

553 To the best of our knowledge, this is the first work that studies the influence  
554 of the constraint subsets change on the features selected by the constraint  
555 scores.

556 Experimental results on three UCI datasets and a face database show that  
557 the proposed score is less sensitive to the constraint changes while selecting  
558 features that provide satisfying classification results.

## 559 **References**

560 Alon, U., Barkai, N., Notterman, D., Gishdagger, K., Ybarradagger, S.,  
561 Mackdagger, D., Levine, A., 1999. Broad patterns of gene expression re-  
562 vealed by clustering analysis of tumor and normal colon tissues probed by  
563 oligonucleotide arrays. *Proceedings of the National Academy of Science of*  
564 *the USA* 96 (12), 745–6750.

565 Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D., 2005. Learning a maha-  
566 lanobis metric from equivalence constraints. *Journal of Machine Learning*  
567 *Research*, 937–965.

568 Bishop, C., Jan. 1996. *Neural Networks for Pattern Recognition*. Oxford  
569 University Press, USA.

- 570 Blake, C., Keogh, E., Merz, C., 1998. UCI repository of machine learning  
571 databases. <http://www.ics.uci.edu/mlearn/MLRepository.html>.
- 572 Dy, J., Brodley, C., 2004. Feature selection for unsupervised learning. *Journal*  
573 *of Machine Learning Research* 5, 845–889.
- 574 Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M.,  
575 Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A.,  
576 Bloomfield, C. D., 1999. Molecular classification of cancer: class discovery  
577 and class prediction by gene expression monitoring. *Science* 286, 531–537.
- 578 Grzegorzewski, P., 2006. The coefficient of concordance for vague data. *Com-*  
579 *putational Statistics and Data Analysis* 51 (1), 314–322.
- 580 He, X., Cai, D., Niyogi, P., Dec. 2005. Laplacian score for feature selection.  
581 In: *Proceedings of the Advances in Neural Information Processing Systems*  
582 *(‘NIPS 05’)*. Vancouver, British Columbia, Canada.
- 583 Liu, H., Motoda, H., Aug. 1998. *Feature extraction construction and selection*  
584 *a data mining perspective*, 1st Edition. Springer.
- 585 Ng, A., Jordan, M., Weiss, Y., Dec. 2001. On spectral clustering: Analysis

- 586 and an algorithm. In: Proceedings of the Advances in Neural Information  
587 Processing Systems ('NIPS 01'). Canada, pp. 849–856.
- 588 Samaria, F., Hartert, A., 1994. Parameterisation of a stochastic model for  
589 human face identification. In: Proceedings of the Second IEEE Workshop  
590 on Applications of Computer Vision 'ACV 94'. Sarasota, Florida, pp. 138–  
591 142.
- 592 Siegel, S., Castellan, N., 1988. Nonparametric statistics for the behavioral  
593 sciences, 2nd Edition. McGraw-Hill, New York.
- 594 Suna, D., Zhang, D., Jun. 2010. Bagging constraint score for feature selection  
595 with pairwise constraints. *Pattern Recognition* 43, 2106–2118.
- 596 von Luxburg., U., 2007. A tutorial on spectral clustering. *Statistics and Com-*  
597 *puting* 17 (4), 395–416.
- 598 Yu, L., Liu, H., 2003. Feature selection for high-dimensional data: A fast  
599 correlation-based filter solution. Vol. 6. pp. 856–863.
- 600 Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and  
601 redundancy. *Journal of Machine Learning Research* 5, 1205–1224.
- 602 Zelink-Manor, L., Perona, P., Dec. 2005. Self-tuning spectral clustering. In:

- 603 Proceedings of the Advances in Neural Information Processing Systems  
604 'ANIPS 05'. Vancouver, British Columbia, Canada, pp. 1601–1608.
- 605 Zhang, D., Chen, S., Zhou, Z., Oct. 2008. Constraint score: A new filter  
606 method for feature selection with pairwise constraints. *Pattern Recognition*  
607 *(41)*, 1440–1451.
- 608 Zhao, J., Lu, K., He, X., 2008. Locality sensitive semi-supervised feature  
609 selection. *Neurocomputing* 71 (10-12), 1842–1849.
- 610 Zhao, Z., Liu, H., Apr. 2007a. Semi-supervised feature selection via spectral  
611 analysis. In: *Proceedings of the SIAM International Conference on Data*  
612 *Mining 'ICDM 07'*. Minneapolis.
- 613 Zhao, Z., Liu, H., 2007b. Semi-supervised feature selection via spectral anal-  
614 ysis. Tech. rep., Computer Science and Engineering (CSE) Department,  
615 Ari-zona State University (ASU).
- 616 Zhao, Z., Liu, H., Aug. 2007c. Spectral feature selection for supervised and  
617 unsupervised learning. In: *Proceedings of the 24th international conference*  
618 *on Machine learning 'ICML 07'*. ACM, Corvalis, Oregon, pp. 1151–1157.