Neurocomputing 121 (2013) 5–13

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A graph Laplacian based approach to semi-supervised feature selection for regression problems

Gauthier Doquire *,[1], Michel Verleysen

Machine Learning Group - ICTEAM, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Available online 24 February 2013

Keywords:
Feature selection
Semi-supervised learning
Graph Laplacian

ABSTRACT

Feature selection is a task of fundamental importance for many data mining or machine learning applications, including regression. Surprisingly, most of the existing feature selection algorithms assume the problems to address are either supervised or unsupervised, while supervised and unsupervised samples are often simultaneously available in real-world applications. Semi-supervised feature selection methods are thus necessary, and many solutions have been proposed recently. However, almost all of them exclusively tackle classification problems. This paper introduces a semi-supervised feature selection algorithm which is specifically designed for regression problems. It relies on the notion of Laplacian score, a quantity recently introduced in the unsupervised framework. Experimental results demonstrate the efficiency of the proposed algorithm.

1. Introduction

When dealing with high-dimensional data sets, feature selection is a step of major importance for many pattern recognition applications, including regression. Indeed, learning with high-dimensional data is generally a complicated task due to many undesirable facts denoted by the term *curse of dimensionality* [1,2].

Moreover, it is quite usual that some features in a data set are either redundant or even totally uninformative; they can decrease the performances of the learning algorithms and make them prone to overfitting [3]. In addition, removing such useless features generally helps reducing the learning time of the prediction models. Eventually, the interpretation of the models and the understanding of the original problem can also benefit from feature selection.

Other dimensionality reduction approaches such as feature extraction or projection [4,5] can as well be efficient to help managing high-dimensional datasets; however, they do not preserve the original features and thus prevent from interpreting the new low-dimensional features. This can be a major drawback in many areas such as medicine or industry, where interpretation is crucial.

Traditional feature selection algorithms are said to be *supervised*, in the sense that the knowledge of the output associated with each training sample (a class label for classification problems or a continuous value for regression ones) is assumed to be available and can be used. On the contrary, *unsupervised* feature selection methods completely ignore the outputs and thus only consider the dataset in order to determine the relevant features. The most obvious example of unsupervised method is simply the evaluation of each feature variance, which gives an indication about the features predictive power.

Halfway between those two situations, in many real-world problems it is easy to obtain a large number of (unsupervised) samples, while only a few labeled samples are generally available. Such a limitation is mainly due to the cost (in terms of time and/or money) needed to obtain the labels. As an example, in medical diagnosis problems, a human expertise is usually required to determine whether or not a patient is ill, based on analyses or radiographies. Such a task can be hard to perform and time-consuming, even for a trained practician. In another field, one may be interested in predicting the fat or sugar content of a sample food based on a spectroscopic analysis. Even if the spectroscopic data can be quickly and easily obtained, getting the true fat or sugar content values will often necessitate destructive tests (such as burning a piece of meat). Obviously, such tests cannot be performed easily on thousands of samples.

The above discussion naturally justifies the development of *semi-supervised* learning, where the few available labels are used to improve learning algorithms that are mainly based on the unsupervised part of the data [6,7]. In this context, many feature selection algorithms have been proposed recently, whose large

* Corresponding author. Tel.: +32 10 47 81 33.
E-mail addresses: gauthier.doquire@uclouvain.be (G. Doquire), michel.verleysen@uclouvain.be (M. Verleysen).
URL: http://www.ucl.ac.be/mlg/ (M. Verleysen).
[1] Gauthier Doquire is funded by a Belgian FRIA grant.

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.10.028

majority are designed to handle classification problems. Indeed, to the best of our knowledge, no work has been done up to now to develop semi-supervised feature selection algorithms specific to regression problems.

This paper addresses this issue by first introducing a supervised feature selection algorithm, which is then extended to achieve semi-supervised feature selection. Both algorithms are specifically designed to handle continuous outputs and are based on the unsupervised developments introduced in [8]. Within the unsupervised framework, the algorithm in [8] scores features according to their locality preserving power. Roughly speaking, good features have close values for close samples and thus preserve the local structure of the data set. In this work, the idea is extended by using distance information between the output of supervised samples. This paper extends preliminary results introduced in [9].

The rest of the paper is organized as follows. Section 2 briefly mentions related works on feature selection and semi-supervised learning. Section 3 describes the original unsupervised Laplacian score. Section 4 introduces the supervised feature selection criterion, while Section 5 presents the semi-supervised algorithm which combines in a simple way the information from supervised and unsupervised samples. Its efficiency is experimentally demonstrated in Section 6. Eventually, Section 7 concludes the work and gives possible directions for further investigations.

2. Related work

Due to its importance, feature selection is a problem that has been widely studied. In the literature, three main approaches can be distinguished. First, wrappers [10] select a subset of features in order to directly maximize the performances of a particular prediction model. They thus require building many of these models and can be very time-consuming in practice. However, they traditionally lead to good prediction performances.

On the other hand, filters look for features maximizing a criterion which does not rely on any specific prediction model. Among possible criteria, the most popular ones are the mutual information [11,12] and the correlation coefficient [13]. Filters are traditionally faster and more generic than wrappers, in the sense that they can be used prior to any prediction model. Eventually, embedded methods, whose most well-known examples are LASSO and its extensions [14,15], perform prediction and feature selection simultaneously by somehow regularizing an objective function.

Even if most of existing feature selection algorithms are designed for supervised problems, feature selection has also been tackled in the unsupervised learning context; Mitra et al. proposed an approach using feature similarity [16] while [17] is based on expectation-maximization clustering. Madsen et al. [18] also proposed to use the dependency between features while a Graph Laplacian based ranking method has been developed in [8]. This last work is at the basis of the present paper and will be presented in more details later. It has also inspired [19], which considers another form of weak supervision: pairwise constraints. More precisely, in the latter context, the exact class labels are not available but it is known, for a limited number of samples, whether or not they belong to same class.

Concerning the semi-supervised paradigm we are interested in, different solutions for feature selection have been proposed. Among many others, Zhao and Liu proposed an approach using spectral analysis [20] while Quinzan et al. introduced an algorithm based on feature clustering, conditional mutual information and conditional entropy [21]. In [22], the authors use the notion of Markov blanket. In [23], the authors propose to solve a class margin maximization problem involving a manifold regularization term. In [24], Zhong et al. introduce a "hybrid" method that selects an initial set of features using the supervised samples, before expending and correcting this set using the unsupervised samples and label propagation techniques. Eventually, in [25], the concept of graph Laplacian is also exploited, with proximity matrices defined through the available class memberships. The common feature of all these works is that they are designed for classification problems only, with no obvious way of extending them to regression problems.

3. Laplacian score

This section describes the Laplacian score (LS), introduced by He et al. [8] in the unsupervised framework. Generally speaking, the method ranks the features according to their locality preserving power, i.e. according to how well they preserve the local structure of the data set.

Consider a data set X. Let f_{ri} denote the rth feature of the ith sample ($i=1\dots m$), $\mathbf{x_i}$ the ith data point and $\mathbf{f_r}$ the rth feature. Build then a similarity graph with m nodes (one for each data point), which contains an edge between node i and node j if the corresponding samples $\mathbf{x_i}$ and $\mathbf{x_j}$ are close, i.e. if $\mathbf{x_i}$ is among the k nearest neighbors of $\mathbf{x_j}$ or conversely. Even if any measure of similarity between the samples can be used, the Euclidean distance will be considered throughout this paper to determine the nearest neighbors of each point.

From the proximity graph, a matrix S^{uns} can be built by setting

$$S_{i,j}^{uns}=\begin{cases} e^{-\|\mathbf{x_i}-\mathbf{x_j}\|_2^2/t} & \text{if } \mathbf{x_i} \text{ and } \mathbf{x_j} \text{ are close,}\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

where t is a suitable positive constant. Define then $D^{uns}=diag(S^{uns}\mathbf{1})$, with $\mathbf{1}=[1\dots 1]^T$, and the graph Laplacian $L^{uns}=D^{uns}-S^{uns}$ [26].

The mean of each feature, weighted by the local density of data points, is then removed: the new features are called $\tilde{\mathbf{f}}_\mathbf{r}$ and are given by $\tilde{\mathbf{f}}_\mathbf{r}=\mathbf{f_r}-(\mathbf{f_r^T}D^{uns}\mathbf{1}/\mathbf{1^T}D^{uns}\mathbf{1})\mathbf{1}$. Using a weighted normalization has a natural interpretation in terms of spectral graph theory [8,26]; however, traditional normalization can as well be used, as suggested in [8]. See also at the end of this section for a more detailed argument for the normalization.

Eventually the Laplacian score of each feature $\mathbf{f_r}$ is computed as

$$L_r=\frac{\tilde{\mathbf{f}}_\mathbf{r}^T L^{uns}\tilde{\mathbf{f}}_\mathbf{r}}{\tilde{\mathbf{f}}_\mathbf{r}^T D^{uns}\tilde{\mathbf{f}}_\mathbf{r}}. \tag{2}$$

Features are ranked according to this score, in increasing order. In Ref. [8], the authors also derive a connection between the LS (2) and the well-known Fisher criterion. As will be made clear in Section 4.2, the numerator of this criterion has a sound interpretation in terms of the ability of features to preserve a given local structure. The denominator of (2) is the weighted variance of the feature $\mathbf{f_r}$, considered as an indicator of the feature predictive power.

We end this section with a brief comment on the reasons why the normalization $\tilde{\mathbf{f}}_\mathbf{r}=\mathbf{f_r}-(\mathbf{f_r^T}D^{uns}\mathbf{1}/\mathbf{1^T}D^{uns}\mathbf{1})\mathbf{1}$ is necessary. The goal in removing the weighted mean is actually to prevent a non-zero constant vector such as $\mathbf{1}$ to be assigned a zero Laplacian score. This is important since a low Laplacian score corresponds to a relevant feature while a constant feature obviously does not contain any information. As detailed in [8], after the proposed normalization, the numerator of (2) cannot be zero for a constant feature if the denominator is not zero. If both quantities are zero, the score of the feature would thus be 0/0, which is an

indetermination; the feature would then be removed from the problem. On the contrary, without the normalization, a constant vector could be assigned a zero score and consequently be ranked as the most relevant feature.

4. Supervised Laplacian score

This section introduces a new supervised feature selection method, based on the Laplacian score (2). This criterion will be useful to eventually propose the semi-supervised algorithm. However, it presents by itself interesting properties for feature selection, as will be illustrated below.

4.1. Definitions

Consider again the training set X containing m samples $\mathbf{x_i}$ described by n features. As we are here concerned with supervised regression problems, an output vector $Y=[y_1 \dots y_m] \in \Re^m$ is also available. If the output Y can reasonably be assumed to be a continuous and smooth function of X, it is quite natural to expect close samples $\mathbf{x_i}$ and $\mathbf{x_j}$ to have close output values y_i and y_j. In that sense, good features are thus expected to have close values for data points whose outputs are close too. Using this idea, a feature selection criterion can be constructed as follows.

Let the matrix S^{sup} be defined as

$$S_{i,j}^{sup}=\begin{cases} e^{-(y_i-y_j)^2/t} & \text{if } x_i \text{ and } x_j \text{ are close,} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and $D^{sup}=diag(S^{sup}\mathbf{1})$, $L^{sup}=D^{sup}-S^{sup}$, $\tilde{\mathbf{f}}_\mathbf{r}=\mathbf{f_r}-(\mathbf{f_r^T}D^{sup}\mathbf{1}/\mathbf{1^T}D^{sup}\mathbf{1})\mathbf{1}$. For the construction of S^{sup}, two points x_i and x_j are considered as close if one of the corresponding outputs (y_i or y_j) is among the k nearest neighbors of the other; t is a suitable (positive) constant. Criterion (2) can again be used to rank features by computing a quantity we call the supervised Laplacian score (SLS)

$$SLS_r=\frac{\tilde{\mathbf{f}}_\mathbf{r}^T L^{sup}\tilde{\mathbf{f}}_\mathbf{r}}{\tilde{\mathbf{f}}_\mathbf{r}^T D^{sup}\tilde{\mathbf{f}}_\mathbf{r}}. \tag{4}$$

4.2. Justification

As stated above, a good feature can be expected to have similar values for points whose outputs are similar too. Consequently, a quite natural objective to minimize for a relevant feature is the following one:

$$\min_{\mathbf{f}} \sum_i \sum_j (f_{ri}-f_{rj})^2 S_{ij}^{sup}. \tag{5}$$

Indeed, if S_{ij}^{sup} becomes large, $(f_{ri}-f_{rj})^2$ has to decrease for the criterion (5) to remain small. Features according to which $\mathbf{x_i}$ and $\mathbf{x_j}$ are not close while they are in the sense of S_{ij}^{sup} are thus penalized. This way, the local structure of the data can be preserved. In the following, the connection between criteria (4) and (5) is shown.

From the definition of the diagonal matrix $D^{sup}=diag(S^{sup}\mathbf{1})$, it can be easily deducted that $D_{ii}^{sup}=\sum_j S_{ij}^{sup}$. Remembering that $L^{sup}=D^{sup}-S^{sup}$, some basic calculations give

$$\begin{aligned} \sum_i \sum_j (f_{ri}-f_{rj})^2 S_{ij}^{sup} &= \sum_i \sum_j (f_{ri}^2+f_{rj}^2-2f_{ri}f_{rj})S_{ij}^{sup} \\ &= \sum_i \sum_j f_{ri}^2 S_{ij}^{sup} + \sum_i \sum_j f_{rj}^2 S_{ij}^{sup} \\ &\quad -2\sum_i \sum_j f_{ri}f_{rj}S_{ij}^{sup} \\ &= 2\mathbf{f_r}^T D^{sup}\mathbf{f_r} - 2\mathbf{f_r}^T S^{sup}\mathbf{f_r} \\ &= 2\mathbf{f_r^T}(D^{sup}-S^{sup})\mathbf{f_r} \\ &= 2\mathbf{f_r}^T L^{sup}\mathbf{f_r}. \end{aligned} \tag{6}$$

Therefore it appears clearly that minimizing the numerator $\tilde{\mathbf{f}}_\mathbf{r}^T L^{sup}\tilde{\mathbf{f}}_\mathbf{r}$ of (4) is equivalent to minimizing (5). The denominator $\tilde{\mathbf{f}}_\mathbf{r}^T D^{sup}\tilde{\mathbf{f}}_\mathbf{r}$, already considered in (2), still penalizes features with a low variance, but also prevents a non-zero constant vector to get a zero score.

4.3. Illustration

SLS is now compared to the well-known and widely studied correlation coefficient and to the mutual information (MI) as a criterion for feature selection. In this section, the value of the parameter k in (3) is set to 5 and the MI is estimated as detailed in [27].

4.3.1. Artificial data sets

SLS is first tested on artificial problems for which the relevant features are known in advance. The objective is to show the ability of the method to actually select relevant features and its greater capability to detect non-linear relationships between each feature and the output than the correlation. To this end, three artificial problems are considered.

The first problem has six features $X_1 \dots X_6$ uniformly distributed on [0; 1]. Its output is a linear combination of three features

$$Y_1 = 5X_1 + 7X_2 - 10X_3. \tag{7}$$

The second problem has eight features $X_1 \dots X_8$ uniformly distributed on [0; 1]. Its output is defined as

$$Y_2 = \cos(2\pi X_1 X_2)\ \sin(2\pi X_3 X_4). \tag{8}$$

The last one consists of four features $X_1 \dots X_4$ uniformly distributed on [0; 1]. The output is defined as

$$Y_3 = X_1^2 X_2^{-2}. \tag{9}$$

For the three problems, the sample size is set to 1000 and 1000 datasets are randomly generated. The criterion of comparison between feature selection methods is the percentage of cases for which all the informative features (three, four and two for Y_1, Y_2 and Y_3 respectively) are the best ranked. Table 1 summarizes the results. As can be seen, when the relationship between the features and the output is non-linear, the proposed SLS clearly outperforms the correlation coefficient, which is restricted to linear dependencies, and compares well with the MI. For the linear problem (7), all the three methods give satisfactory results.

4.3.2. Real-world problems

SLS is now tested on real-world data sets. Since in this case the relevant features are not known in advance, a prediction model has to be used to assess the performances of the feature selection criteria. In this paper, a 5-nearest neighbors (5-NN) model is used, both for its simplicity and its sensitivity to irrelevant features. Two data sets are considered, Delve Census and Orange Juice, which will be described in details in Section 6; all the available samples are used for the experiments.

Table 1
Percentage of experiments for which the relevant features are ranked first on three artificial problems.

Criterion	Y_1	Y_2	Y_3
Correlation	100	25	32
SLS	100	93	100
MI	100	100	100

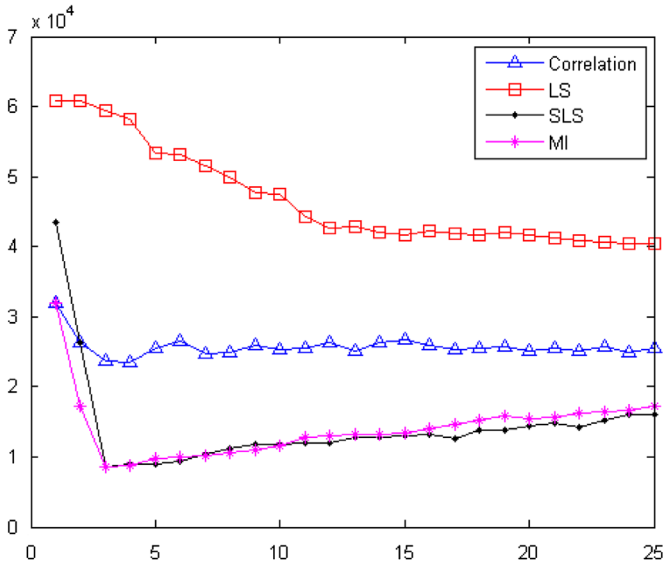

Fig. 1. RMSE of a 5-NN model for two supervised and one unsupervised feature selection criteria on the Delve Census data set.

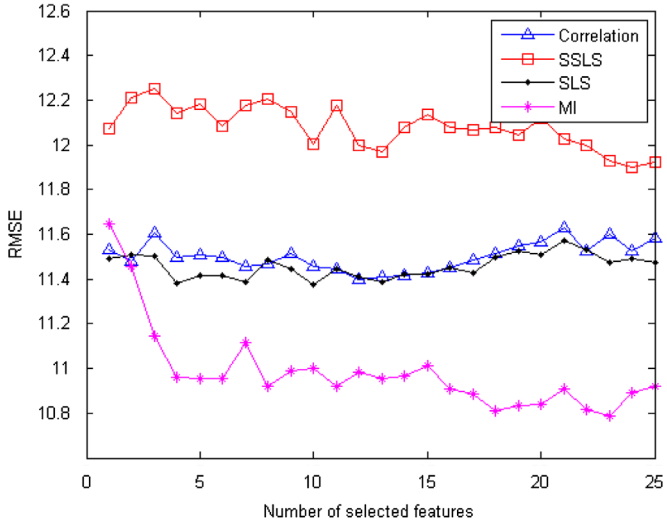

Fig. 2. RMSE of a 5-NN model for two supervised and one unsupervised feature selection criteria on the Orange Juice data set.

Figs. 1 and 2 show the root mean squared error (RMSE) of the 5-NN model as a function of the number of selected features for the two data sets. The RMSE is estimated through a 5-fold cross-validation procedure. For comparison, the results with the unsupervised Laplacian score are also shown. Again, the interest of the SLS against the correlation coefficient can be observed. As expected, the unsupervised method performances are the worse. The MI outperforms its competitors for the Juice dataset and is equivalent to the SLS for the Delve dataset. However, as will be shown in Section 6 (left column of Figs. 3–6), when the number of labeled samples is low, the MI does not perform as well anymore; in this situation, the SLS performances are comparable or slightly better than the ones of the MI. This can be due to the fact that MI is too complex to be reliably evaluated with such a few number of samples.

5. Semi-supervised Laplacian score

When a large number of labeled data points are available, Section 4 shows promising results concerning the use of SLS as a feature selection criterion. However, when the number of labeled samples is small, information from the unlabeled part of the data should also be taken into account.

In this paper, we introduce a semi-supervised feature selection algorithm based on the developments in the two previous sections. As detailed above, both the LS and the SLS are based on the ability of the features to preserve the local structure of the data. In fact, the difference between these two methods precisely lies in the way the local structure is determined. Indeed, the structure is defined from the unsupervised part of data for LS (two points are close if the values of their features are close) and from the output for SLS (two points are close if their outputs are close). In order to combine both pieces of information, a quite intuitive idea is thus to compute the distance between two samples from their outputs if both are known, and from the unsupervised part of the data otherwise. Indeed, Section 4.3 showed that, as could be easily understood, SLS leads to better prediction results than LS and that the supervised information should thus be used preferentially if available. These considerations are the basis of the design of the proposed feature selection criterion.

Let us consider a semi-supervised regression problem, which consists in the training set X and the associated output vector $Y=[y_1 \dots y_s] \in \Re^s$. It is assumed that $s \ll m$, i.e. that the number of supervised samples is small regarding the total number of samples, which is a traditional assumption in semi-supervised learning.

The developments begin by defining the matrix d of pairwise distances between each pair of data points

$$d_{i,j}^2 = \begin{cases} (y_i - y_j)^2 & \text{if } y_i \text{ and } y_j \text{ are known,} \\ \frac{1}{n}\sum_{k=1}^{n}(f_{k,i}-f_{k,j})^2 & \text{otherwise.} \end{cases} \quad (10)$$

In the second case, i.e. if y_i, y_j or both y_i and y_j are unknown, the distance is normalized by the number of features n. This way, all values of d are kept in a comparable range, which would not be the case otherwise, since the dimension of the data set is generally much larger than one. Of course, for this distance normalization to be meaningful, the features $\mathbf{f_r}$ as well as the output vector Y have to be normalized to the same range of values. Details about this normalization are given in Section 6.

Based on d, a matrix S^{semi} is then built as follows:

$$S_{i,j}^{semi} = \begin{cases} e^{-d_{i,j}^2/t} & \text{if } \mathbf{x_i} \text{ and } \mathbf{x_j} \text{ are close} \\ & \text{and } y_i \text{ and/or } y_j \text{ is unknown,} \\ C \times e^{-d_{i,j}^2/t} & \text{if } \mathbf{x_i} \text{ and } \mathbf{x_j} \text{ are close} \\ & \text{and } y_i \text{ and } y_j \text{ are known,} \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

Again, two points are considered as close if one is among the k nearest neighbors of the other one.

As one can notice, a (positive) constant C is introduced to weight the values of $S_{i,j}^{semi}$ corresponding to supervised samples. In practice, this allows us to give more importance to the information coming from the supervised part of the data. Indeed, since the SLS has an obvious advantage over his unsupervised counterpart, it is reasonable to assume the labels to be more important than the unsupervised samples for the feature selection problem.

Similarly to what has been done in the previous section, we then define $D^{semi} = diag(S^{semi}\mathbf{1})$, $L^{semi} = D^{semi} - S^{semi}$ and $\tilde{\mathbf{f}}_\mathbf{r} = \mathbf{f_r} - (\mathbf{f_r^T} D^{semi}\mathbf{1} / \mathbf{1^T} D^{semi}\mathbf{1})\mathbf{1}$.

Eventually, the proposed criterion for semi-supervised feature selection, called semi-supervised Laplacian score (SSLS), can be

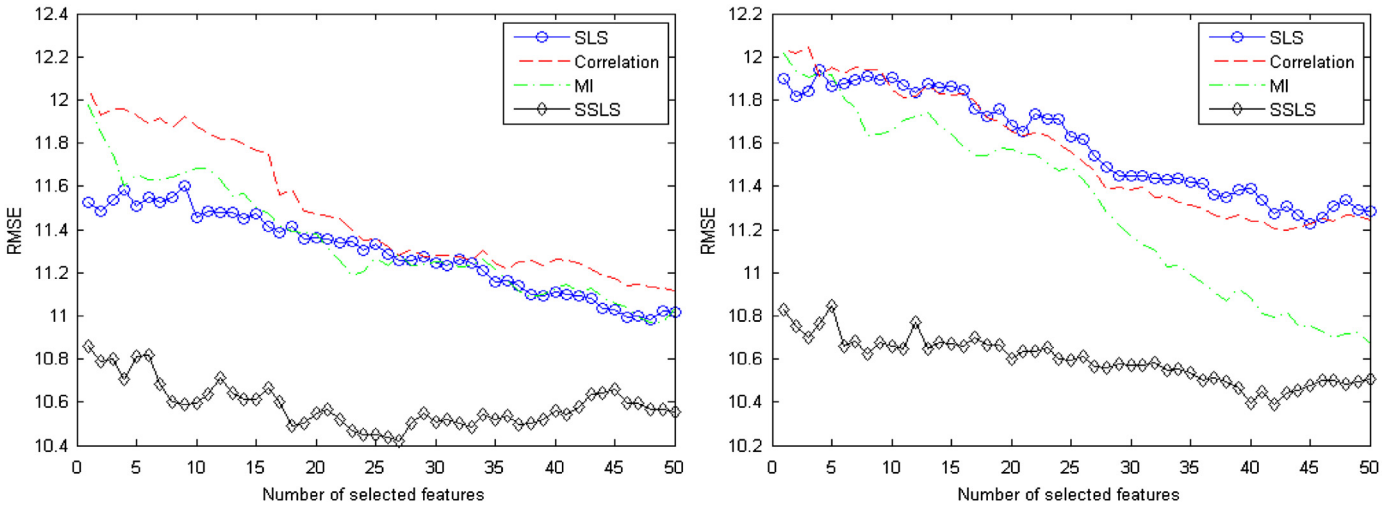

Fig. 3. RMSE of a 5-NN model as a function of the number of selected features with 3% (left) and 5% (right) supervised samples for the Orange Juice data set.

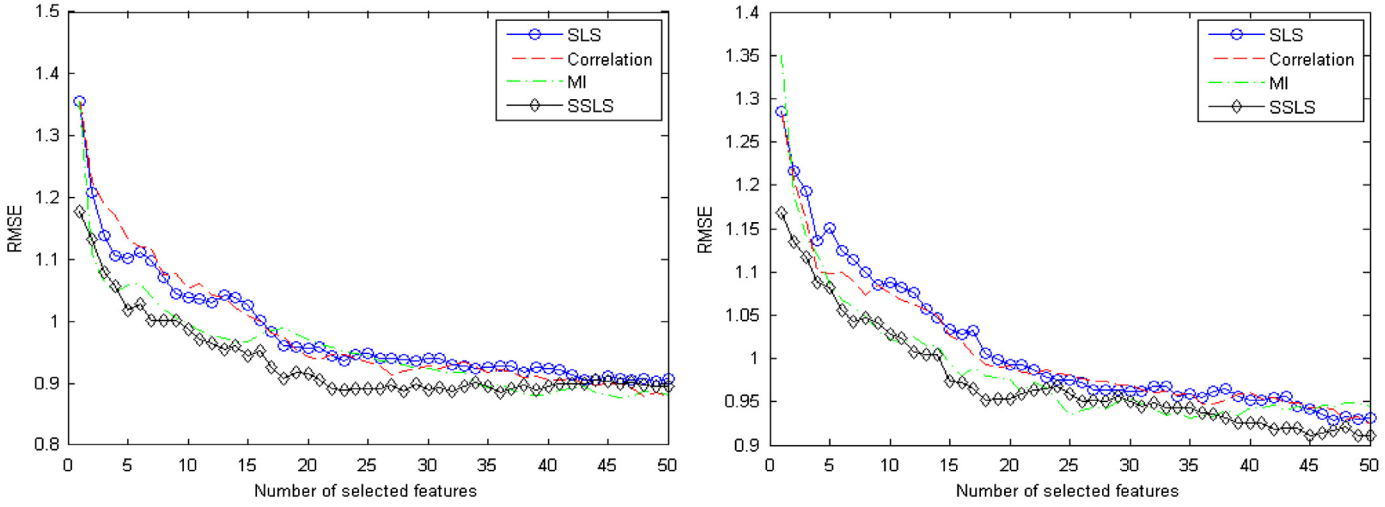

Fig. 4. RMSE of a 5-NN model as a function of the number of selected features with 3% (left) and 5% (right) supervised samples for the Nitrogen data set.

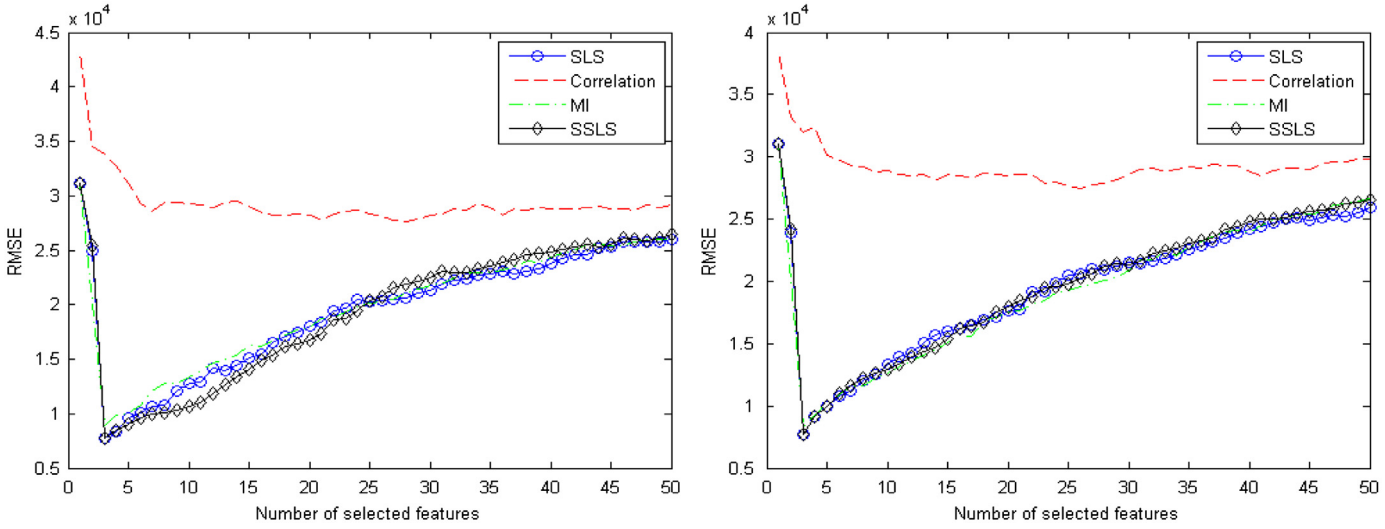

Fig. 5. RMSE of a 5-NN model as a function of the number of selected features with 3% (left) and 5% (right) supervised samples for the Delve Census data set.

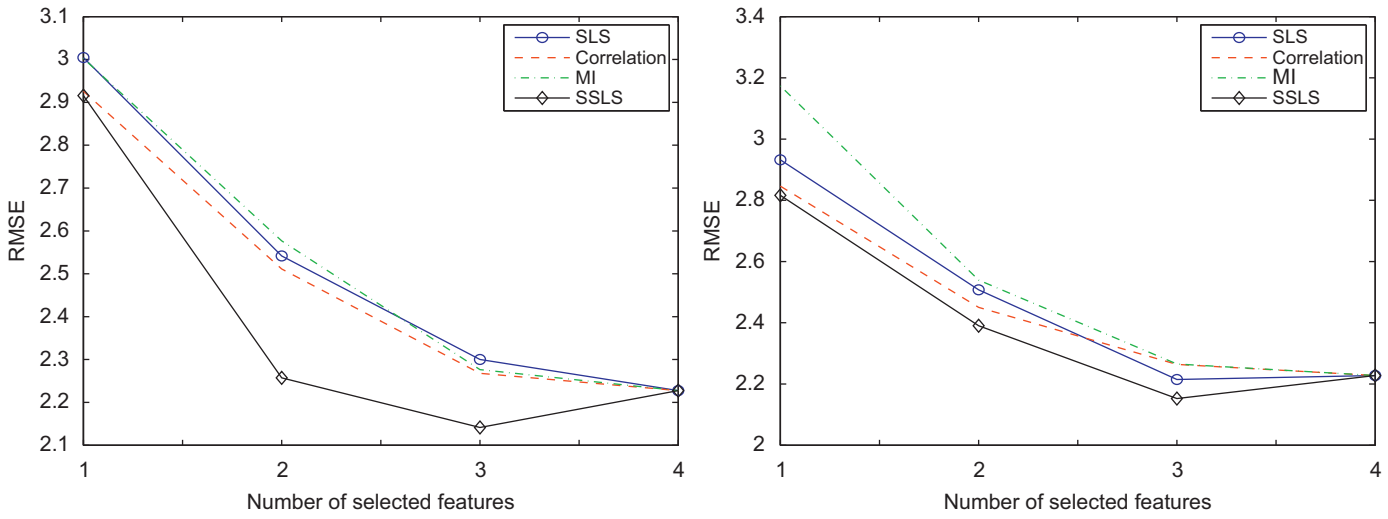

Fig. 6. RMSE of a 5-NN model as a function of the number of selected features with 3% (left) and 5% (right) supervised samples for the Pollen data set.

defined for each feature $\mathbf{f_r}$ as

$$SSLS_r = \frac{\tilde{\mathbf{f}}_\mathbf{r}^T L^{semi} \tilde{\mathbf{f}}_\mathbf{r}}{\tilde{\mathbf{f}}_\mathbf{r}^T D^{semi} \tilde{\mathbf{f}}_\mathbf{r}} \times SLS_r. \tag{12}$$

In this last definition, SLS_r is the supervised Laplacian score introduced in Section 4, computed using the few supervised samples only. Developments similar to (6) can also give a justification to the first term of the proposed criterion.

This criterion (12) thus combines the influence of both the unsupervised and the supervised part of the data, but gives however this last part more importance. More precisely, the SSLS is the SLS corrected with a term slightly influenced by the sole values of X. As will be seen in the next section, even such a small influence can significantly improve the feature selection performances compared to methods using the supervised samples only. In this work, the product has been chosen to combine the information from the supervised and the unsupervised part of the data, as preliminary results suggested the interest of doing so. Obviously, other possibilities could as well be considered.

6. Experimental results and discussion

This section illustrates the interest of the proposed semi-supervised feature selection approach. First, the impact of different feature selection criteria on the performances of a prediction model is studied. Then, it is quickly verified that the proposed SSLS is able to efficiently use the few supervised samples, by showing how its performances are influenced by the number of labeled samples. Eventually, we discuss the influence of the different parameters of our method and give some simple considerations to choose empirically good values for these parameters.

6.1. Prediction performances

Four real-world data sets are used in this section. The first two ones concern near-infrared spectra analysis problems. With the Orange Juice dataset, the goal is to estimate the level of saccharose of orange juice samples from their measured near-infrared spectrum. About 218 samples each defined by 700 features are available. The dataset can be downloaded from the website of the UCL's Machine Learning Group.[2] For the Nitrogen dataset, containing originally 141 spectra discretized at 1050 different wavelengths, the objective is the prediction of the nitrogen content of a grass sample. The dataset is available from the Analytical Spectroscopy Research Group of the University of Kentucky.[3] In order to lower the original number of features, each spectrum is represented by its coordinates in a B-splines basis as a preprocessing step (see [28] for details). About 105 new features are built this way.

The next data set is the well-known Delve-Census, from which only the 2048 first samples are considered. The data can be obtained from the University of Toronto.[4] The objective is to predict the median price of houses in different small regions. Originally, each sample was described by 139 demographic features about these regions; however, only 104 features are considered here, since those which are too correlated with the output have not been kept for the experiments.

Eventually, experiments are also carried out on the Pollen data set from the StatLib repository.[5] It is a synthetic data set, reproducing characteristics of pollen grains. There are 481 samples and four features.

The proposed SSLS is compared with the MI and the correlation coefficient, but also with the SLS, its supervised counterpart. Indeed, it is important to check that, with a small number of supervised samples, supervised methods are not able to perform well. As the results with the unsupervised LS have already been seen as being the worse ones [9], they are not reproduced here, for the sake of clarity.

As in Section 4.3.2, the compared feature selection algorithms are evaluated through the root mean squared error (RMSE) of a 5-NN model. More precisely, the experimental setup is the following one. First, a few supervised samples are randomly selected from the training set. Features are then selected on the training set; all samples of the training set are considered for the SSLS, while only the supervised ones are used to evaluate the MI, the correlation coefficient and the SLS as those three last methods are not able to take unsupervised samples into account. A 5-NN prediction model that considers only the selected features is then

[2] http://www.ucl.ac.be/mlg/.
[3] http://kerouac.pharm.uky.edu/asrg/cnirs/.
[4] http://www.cs.toronto.edu/~delve/data/census-house/desc.html.
[5] http://lib.stat.cmu.edu/datasets/.

used to predict the output of the points of an independent test set. For the prediction step, the samples in the training set are all assumed to be supervised. This is because the model would probably perform badly if a too low number of supervised samples were available; the obtained results would then have no meaning. This assumption actually ensures that the performances reflect the quality of the feature selection itself, and are not too much influenced by the prediction model. The compared algorithms are tested with 3% and 5% of randomly selected supervised samples. The RMSE is estimated through a 5-fold cross validation procedure repeated 10 times.

The parameters are set follows. t is set to 1; five neighbors are considered for computing the unsupervised (1) and supervised (4) score, while 30 neighbors are considered for the semi-supervised scores (11). Indeed, the number of supervised samples being small, increasing the number of neighbors considered in the analysis allows to take such samples into account with a greater probability. The parameter C in (11) is set to 5. This quite moderate value gives the supervised samples a large importance, but nevertheless gives a significant weight to the information coming from the unsupervised data points. The maximum number of selected features is set to 50. Eventually, before any distance computation, the features and the output vector are normalized by removing their mean and dividing them by their standard deviation. For the prediction step, the outputs are not normalized anymore.

The parameter values have been determined empirically, as they consistently led to good performances in the experiments. However, those values should probably depend on the data set and a way to automatically set them would be of great interest.

Figs. 3–6 show the RMSE as a function of the number of selected features for the different data sets. Results first show how the use of unlabeled data can improve the feature selection procedure for regression purposes when compared with a purely supervised approach. Indeed, it is particularly obvious for the Orange Juice (Fig. 3) and the Pollen (Fig. 6) data sets that the SSLS outperforms the SLS when only a few supervised samples are available. Results on the Nitrogen (Fig. 4) data set are also in favor of the semi-supervised criterion, which is more able to quickly detect relevant variables (at least in terms of regression performances). Results obtained with the Delve data set (Fig. 5) are quite comparable for both methods. These observations indicate that the knowledge coming from the unsupervised samples is efficiently taken into account in the feature selection procedure, and that it positively impacts the determination of relevant features.

Moreover, it can also be observed that the proposed SSLS performs better than the correlation coefficient for the four examples. More precisely, with only a few exceptions for the Nitrogen data set and 3% supervised samples, the RMSE obtained with the SSLS is never larger than the one obtained with the correlation coefficient. The SSLS thus seems to have an advantage over both unsupervised and supervised approach to feature selection. The SSLS also outperforms the MI for the Juice and the Pollen datasets. Results are comparable on the two other datasets.

6.2. *Efficient use of the supervised samples*

One desirable property of a semi-supervised algorithm is that it should be able to efficiently use the information coming from the few (and precious) labeled samples. More specifically, it is interesting to see how the performances of the proposed criterion are affected by the number of supervised samples, especially when a very low number of them is available. To this end, the same experimental framework as in Section 6.1 has been considered and the SSLS criterion has been tested with different small numbers of supervised samples.

Figs. 7–9 present the results on three data sets (the results on the second spectroscopic data sets are not shown for concision reasons). As can be seen, the SSLS always benefits from the addition of supervised samples. Indeed, for the three data sets, the performances of the 5-NN prediction model with the selected features obviously increase with the number of supervised samples.

Of course, it can be reasonably supposed that the above conclusions will only be true when a small amount of supervised samples are considered. Indeed, it is likely that, as the number of available supervised samples grows, the behavior of both the SLS and the SSLS will become similar; consequently, the influence of the unsupervised samples in the SSLS will tend to decrease in that case. This is, however, not a drawback of the proposed methodology, since we are particularly interested in the situations where

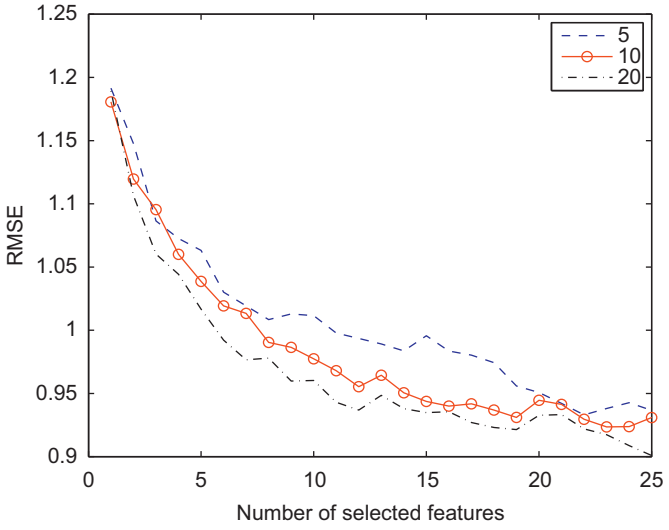

Fig. 7. Performances of a 5-NN model as a function of the number of features selected with the SSLS criterion and different numbers of supervised samples on the Nitrogen data set.

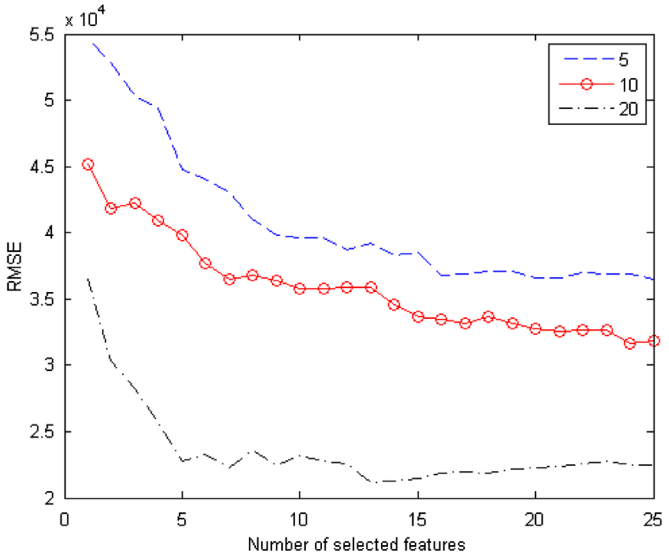

Fig. 8. Performances of a 5-NN model as a function of the number of features selected with the SSLS criterion and different numbers of supervised samples on the Delve Census data set.

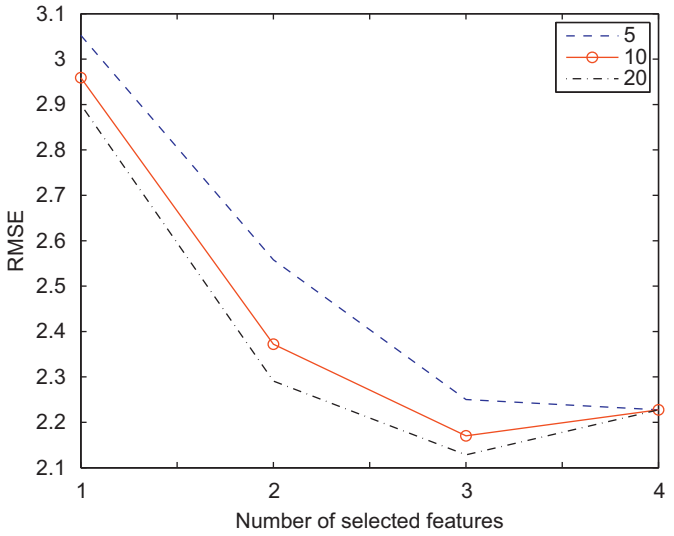

Fig. 9. Performances of a 5-NN model as a function of the number of features selected with the SSLS criterion and different numbers of supervised samples on the Pollen data set.

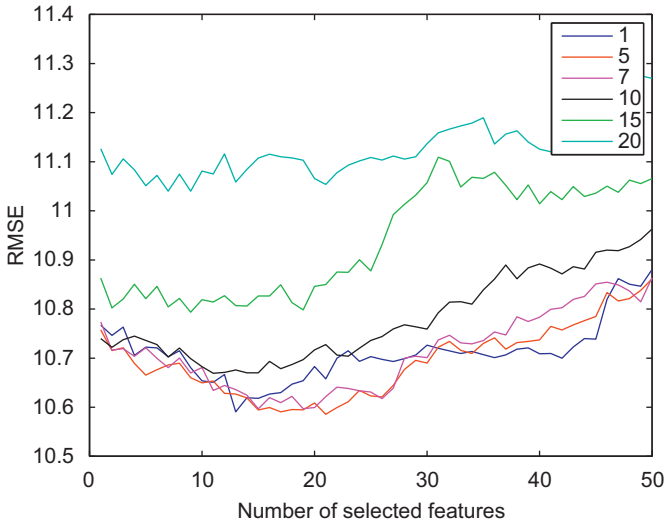

Fig. 10. Performances of a 5-NN model as a function of the number of features selected with the SSLS criterion and different values of the parameter C on the Orange Juice data set. There are 10 supervised samples. Note that the different RMSE scale compared to Fig. 3.

very few supervised data points are available or, said otherwise, where traditional supervised algorithms do not have enough information to perform well.

6.3. Discussion on the parameters

The proposed algorithm involves various parameters whose values have to be determined in practice. In this section, we consequently discuss the actual influence of these parameters and efficient solutions to fix them.

The first parameter is the number of neighbors considered in the construction of the proximity matrices (1), (3) and (11). For the unsupervised and supervised score, a small value of 5 is chosen, in order for the graph to reflect the local structure of the data. This value is successfully used in related works such as [8,25]. For the semi-supervised score (11), the number of neighbors has been increased, to take the supervised samples into account with a greater probability. While the particular value of 30 has been chosen, any intermediate value can in practice be used. Preliminary experiments with 50 neighbors, and up to 100 for the larger Delve dataset, confirmed this intuition with non-significant differences observed in the final performances.

The second parameter is t in Eqs. (1), (3) and (11). It has actually been introduced essentially to remain consistent with the notations of the original LS paper [8]. However, we have chosen to set it at 1 for the following reasons. First, the data we deal with are assumed to be normalized to have zero mean and unit variance. Then, we only consider a small number of neighbors to build the proximity matrix. It is therefore not useful to consider a fast decaying exponential function to represent the local structure of the data. The same choice would not necessarily be adequate if, for instance, all samples were considered to build matrices (1), (3) and (11).

Eventually, parameter C aims at giving more weight to the supervised information in the construction of S^{semi} (11). While this parameter could be helpful in practice, experiments on the four considered datasets have shown that its value only slightly impacts the results. As an example, Fig. 10 shows the performances of the SSLS with 10 supervised samples and different values for C. As expected, when the value of C remains low, the performances of the method are close. When the value of C becomes too high, the unsupervised samples are not taken into account anymore and the performances of the method decrease. Empirically, the value of C can thus be set to 1 to get satisfactory results. When the proportion of supervised samples is extremely low (e.g. less than 1%), we however suggest a moderate value of $C=5$.

In conclusion, the number of neighbors in Eq. (11) is not decisive while parameters t and C can reasonably be omitted in practice. This consequently simplifies the proposed feature selection procedure. We however mention those parameters as they can prove to be useful in some situations, in particular when the total number of samples allows the user to use cross-validation procedures.

7. Conclusions and future work

This paper tackles the important, and surprisingly not studied, problem of feature selection for semi-supervised regression problems. To this end, two algorithms are proposed.

The first one is purely supervised and serves as a basis to the development of the semi-supervised one. Nevertheless, experiments show that the supervised method, called supervised Laplacian score (SLS), presents interesting properties for supervised feature selection, especially when compared with the very popular correlation coefficient criterion. Both proposed algorithms are inspired by the Laplacian score, a feature selection criterion recently introduced in the unsupervised context. Both methods are based on the ability of the features to locally preserve a defined structure of the data. In other words, the proposed criteria give the highest rating to the features which are the most coherent with a similarity measure defined between samples.

In supervised learning, the similarities are estimated using the outputs only as they are supposed to be particularly useful in a regression problem context. For semi-supervised problems, similarities are computed through the outputs if they are known and through the (unlabeled) data points otherwise. This leads to a semi-supervised score, which is then used to transform the SLS into the proposed criterion, called the semi-supervised Laplacian score (SSLS).

Experimental results prove the interest of the proposed approach, for both the supervised and the semi-supervised feature selection

problems. More precisely, for the real-world data sets considered in the paper, SSLS appears to be superior to the correlation coefficient and to the unsupervised Laplacian score. This last observation shows that the proposed SSLS is effectively able to take advantage of the few supervised samples to detect relevant features.

Moreover, SSLS also leads to better prediction performances than its supervised version; it thus also takes advantage of the unsupervised samples when the number of supervised points is particularly low.

In addition, experiments indicate that when few labeled samples are available, the performances of the SSLS grow with the number of these instances. This indicates that the method integrates in an efficient way the information coming from the supervised part of the data, which is a very desirable property for semi-supervised algorithms.

References

[1] R.E. Bellman, Adaptive Control Processes—A Guided Tour, Princeton University Press, Princeton, New Jersey, USA, 1961.
[2] M. Verleysen, Learning high-dimensional data, Limitations Future Trends Neural Comput. 186 (2003) 141–162.
[3] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[4] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, Springer-Verlag New York, Inc., 2006.
[5] J.A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer, New York, London, 2007.
[6] O. Chapelle, B. Schölkopf, A. Zien (Eds.), Semi-Supervised Learning, MIT Press, Cambridge, MA, 2006.
[7] X. Zhu, A.B. Goldberg, Introduction to Semi-Supervised Learning, Morgan & Claypool Publishers, 2009.
[8] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: NIPS, vol. 17, 2005.
[9] G. Doquire, M. Verleysen, Graph Laplacian for semi-supervised feature selection in regression problems, in: IWANN'11, Springer-Verlag, 2011, pp. 248–255.
[10] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[11] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5 (1994) 537–550.
[12] F. Rossi, A. Lendasse, D. Francois, V. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemometrics Intell. Lab. Syst. 80 (2006) 215–226.
[13] M. Hall, Correlation-Based Feature Selection for Machine Learning, Ph.D. Thesis, University of Waikato, 1999.
[14] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc. Ser. B 58 (1) (1996) 267–288.
[15] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. Roy. Stat. Soc. Ser. B 68 (2006) 49–67.
[16] P. Mitra, S. Member, C.A. Murthy, S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 301–312.
[17] J.G. Dy, C.E. Brodley, Feature selection for unsupervised learning, J. Mach. Learn. Res. 5 (2004) 845–889.
[18] N.S. Madsen, C. Thomsen, J.M. Pena, Unsupervised feature subset selection, in: Proceedings of the Workshop on Probabilistic Graphical Models for Classification (PKDD'03), 2003, pp. 71–82.
[19] D. Zhang, S. Chen, Z.-H. Zhou, Constraint score: a new filter method for feature selection with pairwise constraints, Pattern Recognition 41 (2008) 1440–1451.
[20] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of the 24th International Conference on Machine Learning, ICML '07, 2007, pp. 1151–1157.
[21] I. Quinzán, J. M. Sotoca, F. Pla, Clustering-based feature selection in semi-supervised problems, in: ISDA, IEEE Computer Society, 2009, pp. 535–540.
[22] R. Cai, Z. Zhang, Z. Hao, Bassum: a Bayesian semi-supervised method for classification feature selection, Pattern Recognition 44 (4) (2011) 811–820.
[23] Z.L. Xu, I. King, M.R.T. Lyu, R. Jin, Discriminative semi-supervised feature selection via manifold regularization, IEEE Trans. Neural Network. 21 (2010) 1033–1047.
[24] E. Zhong, S. Xie, W. Fan, J. Ren, J. Peng, K. Zhang, Graph-based iterative hybrid feature selection, in: Proceedings of the 2008 8th IEEE International Conference on Data Mining, ICDM '08, 2008, pp. 1133–1138.
[25] J. Zhao, K. Lu, X. He, Locality sensitive semi-supervised feature selection, Neurocomputing 71 (10-12) (2008) 1842–1849.
[26] F.R.K. Chung, Spectral Graph Theory, American Mathematical Society, 1997.
[27] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E. 69 (2004) 066138 http://dx.doi.org/10.1103/PhysRevE.69.066138.
[28] F. Rossi, N. Delannay, B. Conan-Guez, M. Verleysen, Representation of functional data in neural networks, Neurocomputing 64 (2005) 183–210.

Gauthier Doquire was born in 1987 in Belgium. He received the MS in Applied Mathematics from the Université catholique de Louvain (Belgium), in 2009. He is currently a PhD student at the Machine Learning Group of the same university. His research interests include machine learning, feature selection and mutual information estimation.

Michel Verleysen was born in 1965 in Belgium. He received the MS and the PhD degrees in Electrical Engineering from the Université catholique de Louvain (Belgium), in 1987 and 1992 respectively. He was an Invited Professor at the Swiss Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université Paris 1-Panthéon-Sorbonne, in 2002–2004. He is now a Research Director of the Belgian Fonds National de la Recherche Scientifique (FNRS) and Lecturer at the Université catholique de Louvain. He is Editor-in-Chief of the Neural Processing Letters journal, Chairman of the Annual European Symposium on Artificial Neural Networks (ESANN) Conference, Associate Editor of the IEEE Transactions on Neural Networks journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He is the author or the co-author of about 200 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the Co-author of the scientific popularization book on artificial neural networks in the series "Que Sais-Je?", in French. His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, non-linear statistics, adaptive signal processing, and high-dimensional data analysis.