Published in Proceedings of 14th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2017, 2017, 30 Oct.-3 Nov., Hammamet, Tunisia, which should be cited to refer to this work.

Semi supervised relevance learning for feature selection on high dimensional data

Afef Ben Brahim [1]
Email: afef.benbrahim@yahoo.fr
Alexandros Kalousis[2,3]
Email: Alexandros.Kalousis@unige.ch
[1] Université de Tunis, Tunis Business School, Tunisia
[2]Geneva School of Business administration, HES-SO University of Applied Sciences of Western Switzerland
[3]University of Geneva, Switzerland

***Abstract*—Nowadays, the advanced technologies make amounts of data growing in a fast paced way. In many application fields, this trend concerns specially dimensions of the data. It is the case where features are about thousands and tens of thousands, while the number of instances is much smaller. This phenomenon is known as the curse of dimensionality and it results in modest classification performance and feature selection instability. In order to deal with this issue, we propose a new feature selection approach that makes use of background knowledge about some dimensions known to be more relevant, as a means of directing the feature selection process. In this approach, prior knowledge about some features is used to learn new relevant features by a semi supervised approach. Experiments on three high dimensional data sets show promising results on both classification performance and stability of feature selection.**

I. Introduction

The rapid technological developments in different life domains increase the amounts of data characterized by large number of features. Unfortunately, standard machine learning methods are not designed to handle such data setting where the number of samples is small while the number of features associated with the raw data is in the order of thousands or tens of thousands. Feature solution can be a solution to deal with this issue as it reduces data dimensionality by removing irrelevant and redundant features [1]. However, the curse of dimensionality phenomenon has also a negative impact on stability of feature selection which is defined as the sensitivity of a method to variations in the training set [2].

Feature selection techniques often use an evaluation function that measure the relevance of the features to the prediction (filter model) or on the performance of a specific predictor (wrapper model) [3]. Filters use only properties of the data to select features, thus they produce a feature set which is not tuned to a specific type of predictive model. Filters yield an explicit best feature subset or feature's ranking by assigning a score to each feature independently. Mutual information is among the most known measures to rank features [4]. Ranking methods ignore redundancy and inevitably fail in situations where only a combined set of features is predictive of the target function. However, they are usually fast and useful in most real-world problems.

Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

Wrapper methods use the error rate of a predictive model to score a feature subset. This score is obtained by training a specific classification model built with a given feature subset and testing it on the hold-out set. Training a new model for each subset makes wrappers very computationally intensive, but usually provide the best performing feature set for that particular classification model. Recursive Feature Elimination (RFE) [5] is a wrapper selection method for linear Support Vector Machine (SVM). In each round it measures the quality of candidate features by training SVM and eliminating features with the lowest weights.

In most feature selection methods, it is usually assumed that all features are equally relevant before the selection procedure. However, some knowledge about the relevance of a fraction of the features can be available in many areas. This can be considered as a partial supervision on the dimensions of a feature selection procedure given that the available knowledge concerns only a fraction of the features. Traditional feature selection algorithms ignore this type of prior knowledge.

In this paper, we propose a robust wrapper feature selection method based on prior knowledge. This method makes use of a partial supervision on some features assumed a priori to be more relevant. Prior knowledge about these dimensions known to be more relevant is incorporated as a means of guiding the feature selection process. Iteratively we make use of the initial prior knowledge and the previously selected features to expand a subset of highly relevant features in a pre-processing phase of feature selection.

The reminder of the paper is organized as follows. Section 2 presents the concept of feature selection by incorporating prior knowledge. We describe our proposed feature selection approach based on prior knowledge in Section 3. In Section

4, we conduct an experimental study on three high dimensional data sets and with comparison to two feature selection techniques. Section 5 concludes this paper.

II. Feature selection using prior knowledge

The availability of prior knowledge about how features can be related to the prediction task will always help feature selection and its subsequent application. This is the case for example when the biological relevance of features can be ascertained. In this case, potentially relevant features can be favored and by another hand, irrelevant ones can be eliminated. Prior knowledge is any information about features that can be used in feature selection to guide the selection process. It is either obtained from domain experts, relevant publications or extracted from relevant data sets via transfer learning [6]. The integration of prior knowledge in the feature selection process can improve the obtained selection result and thus improve the classification performance. Some studies in the literature have explored this direction of incorporating prior knowledge in feature selection. In the context of SVM, [7] proposed a framework that incorporates prior knowledge on features, represented by meta-features, into the learning process. They assume that a weight is assigned to each feature, as in linear discrimination, and they use the meta-features to define a prior on the weights. This prior is based on a Gaussian process and the weights are assumed to be a smooth function of the meta-features.

In [8], authors extended three existing feature selection methods by incorporating prior knowledge about some dimensions known to be more relevant in order to improve the selection stability and the classication performance. They compared them with their original versions, which do not integrate prior knowledge, and showed that integrating prior knowledge increased stability in most cases compared to classical approaches.

Taskar et al. used meta-features of words for text classification when there are features (words) that are unseen in the training set, but appear in the test set. In their work, features are words and meta-features are words in the neighborhood of each word. They used meta features to predict the role of words that are unseen in the training set [9]. Other ideas using feature properties to produce or select good features can be found in the literature and have been applied in various applications. In [10], Lee et al. used transfer learning to construct an informative prior on feature relevance. They assumed that features themselves have meta-features that are predictive of their relevance to the prediction task. They, modeled their relevance as a function of the meta-features using hyperparameters called meta-priors which are learned from an ensemble of related prediction tasks sharing a similar relevance structure.

In [11], authors proposed the partially-supervized-l2-approximation to zero-norm minimization algorithm (PS-l2-AROM) to integrate prior knowledge about some genes known as clinical markers to discriminate DLBCL tissues from Follicular Lymphomas. Before the feature selection process, they assign a relevance value for those genes assumed to be more relevant. The PS-AROM methods modify a linear model objective function, called l1-AROM [12], by adding a prior relevance vector $\beta = [\beta_1, ..., \beta_d]$ defined over the input dimensions. The optimization problem of PS-l2-AROM penalizes the least those dimensions which are assumed a priori more relevant and thus guide the feature selection process. Iteratively, an objective function is solved given the previous features weight vector w along with the fixed relevance vector β, and the process is iterated till convergence. The original l2-AROM method is obtained when $\beta_j = 1$, $\forall$ feature a_j, in other words, without prior preference between input features.

While in [11] the feature selection algorithm is modified by integrating prior knowledge only one time in the feature selection process, in our proposed approach prior knowledge is expanded and integrated iteratively into the feature selection algorithm. Our formulation encompasses a more advanced framework which takes advantage of prior knowledge to search in a first step for more relevant features based only on their neighborhood with features assumed a priori relevant, then in a second step use the extended set of a priori relevant features to be integrated in the feature selection which will give the final feature subset.

III. Proposed approach: Semi-Supervised-l2AROM (SS-L2AROM)

The proposed feature selection framework, which we called SS-L2AROM, consists of two phases. The aim of the first phase is to learn new relevant features by a semi supervised approach. The extended subset of relevant features will be used as prior knowledge to be integrated into the second step to guide the feature selection process. This interactive process is iterated until an optimal number of features is obtained. We implement a feature selection algorithm based on the proposed approach.

Let X be a matrix containing m instances $\mathbf{x}_i = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d$, where d is the number of features, and $\mathbf{y}_i = (y_1, \ldots, y_m), i = 1, .., m$ the vector of class labels for the m instances. Let A be the set of features $\mathbf{a}_j = (a_1, \ldots, a_d), j = 1, .., d$ where $d >> m$. We denote by $R_n \subseteq A$ the set of features that are known to be relevant based on prior knowledge at iteration n. $\beta = [\beta_1, ..., \beta_d]$ is a vector of background knowledge about the input dimensions, the higher the value of β_j the more relevant the corresponding feature is a priori assumed.

Our proposed approach consists initially of solving a semi supervised problem where the training set is given by $\mathbf{X}'$ the transpose of $\mathbf{X}$, i.e the $j^t h$ row is $a_j = (a_{j1}, .., a_{jm})$. The feature a_j will be labeled as Relevant (1) if $a_j \in R_n$ and Unknown (0) otherwise. After this step, and to extend the set of relevant features, an additional set of features predicted as relevant P_n is obtained and added to R_n such that $R'_n = R_n \cup P_n$. The second step of our approach consists of applying on the original matrix $\mathbf{X}$ a feature selection algorithm that can handle prior knowledge on feature relevance using R'_n as the set of a priori relevant features. This step yields a new selected

Algorithm 1 Semi-supervised relevance learning

```
Input:
$[\mathbf{X}^T, \beta_n, R_n]$

$R'_n = R_n$
$\forall a_j \in \overline{R_n}$
    $\mathbf{S}_{a_j}$ = k-nn($a_j$)
if $\forall g \in \mathbf{S}_{a_j}, g$ is relevant then
        $R'_n := R'_n \cup a_j$, i.e. $a_j$ is relevant
    $\beta'_n$=Update($[\beta_n, R'_n]$)
end if
return $R'_n, \beta'_n$
```

feature set denoted by R_{n+1}. This process is iterated until the number of final selected features reaches a desired feature set cardinality. The proposed algorithm is summarized below.

A. Algorithm

The two main steps of our algorithm are detailed in the following.

1) Pre-processing: For the purpose of our semi supervised problem, aiming at predicting new relevant features using a priori relevant ones, we proceed with data transformation in order to make it fit with the problem. Initially, we take the transpose of the data matrix $\mathbf{X}$ in such way that features become the training instances. Then, each feature is assigned a label indicating whether it is a priori relevant (Relevant (1)) or not (Unknown (0)).

2) First phase: Semi-supervised relevance learning: In the first stage of the ith iteration we solve a semi-supervised problem where we are given a vector, β_n, which describes whether a feature is known to be relevant or not, to find additional relevant features if they exist. This is conducted by applying some semi-supervised algorithm which returns an updated feature relevance vector β'_n.

Now, using a k-nearest neighbor algorithm, distances are calculated between a priori relevant features and the remaining features. For each feature a_j, which is part of the features which we do not know whether they are relevant or not, i.e. $a_j \in \overline{R_n}$, we need to find its k nearest neighbors, $\mathbf{S}_{a_j}$. If all its nearest neighbors are known to be relevant then we denote also a_j as relevant. It is very important that the semi-supervised algorithm is well-behaved, i.e. it will not continue producing relevant features in a trivial way until we get the full feature set. However, it will stop at some point. The s nearest neighbors to the a priori relevant features are chosen to extend R_n to R'_n. The vector of prior knowledge β_n is updated to β'_n based on the a priori relevant feature subset R'_n such that each component of this vector is assigned a value of 10 if a feature $\mathbf{a}_j \in R'_n$, and a value of 1 otherwise. The algorithm for the first phase is given in Algorithm 1.

3) Second phase: Application of feature selection algorithm: In the next stage of the algorithm we go back to the original data matrix $\mathbf{X}$ and apply a feature selection algorithm to handle information about prior feature relevance which will be presented to the selection algorithm as weights vector. As discussed before PS-l2-AROM [11] is an example of such a method on which prior knowledge should be also a vector of weights. We consider it as feature selection technique for the second phase of our approach.

Algorithm 2 Feature Selection with Background Knowledge

```
Input:
$\mathbf{X}$: an $m \times d$ dataset
$\mathbf{y}$: $m$-length vector of class labels
$R_0$: set of a priori relevant features.
$\beta_0$: $d$-length vector characterizing features as a-priori (10)
relevant or not-known (0)
$p$: percentage of additional features to include in each step
of the iteration at the feature selection step.
$\epsilon$: tolerance variable determining when the algorithm converged; should be set to a small value, e.g. 0.01.
$n = 0$
$R_n = R_0$
$\beta_n = \beta_0$
repeat
    $[R'_n, \beta'_n]$ = SemiSup($[\mathbf{X}^T, \beta_n, R_n]$)
    $k = (1 + p) \times |R'_n|$ (number of features to select)
    $R_{n+1}$ = PS-l2-AROM($[\mathbf{X}, \mathbf{y}], \beta'_n, k$)
    $n = n + 1$
until $\frac{|R_n \cap R_{n+1}|}{|R_n \cup R_{n+1}|} \leq \epsilon$
```

PS-l2-AROM algorithm is applied in the second step of our algorithm. Iteratively the minimization problem in PS-l2-AROM algorithm is solved given the relevance vector β_n obtained in the first step. Iterations terminate when there are no important differences between the features indicated as relevant in step n by the vector β_n and the ones indicated as relevant in the step $n + 1$ by β_{n+1}.

A crucial point here is whether there is a monotonic increase in R_n vector, i.e. as we move from step n to $n + 1$ do we always have $R_n \subseteq R_{n+1}$? This obviously depends on the behavior of the semi-supervised learning and the feature selection algorithm that we have selected for stages one and two. So we need to study the convergence behavior of the two-step algorithm. This means that we should trace the quantity $\frac{|R_n \cap R_{n+1}|}{|R_n \cup R_{n+1}|}$ as a function of n.

At the semi-supervised step we retrieve $|R'_n|$ features. Then, at the feature selection step we allow the feature selection algorithm to select at least as many features as possible such that $|R_{n+1}| \geq |R'_n|$. In order to control the number of features between the semi-supervised step and the feature selection step, we set the number of features to select as follows: $|R_{n+1}| = (1 + p) \times |R'_n|$ i.e. the number of features that are retained in the feature selection step should be as many as the ones in R'_n plus one small percentage, p. The algorithm is described in Algorithm 2.

Another important point is to study the convergence properties of the algorithm. Basically, this means that at some point the semi-supervised algorithm does not produce anymore

Dataset	No. samples	No. features	No. a priori relevant features
Bladder cancer	31	3036	11
DLBCL	77	7029	2
Lung cancer	181	12533	8

TABLE I: Datasets characteristics

additional relevant features, or produces very few ones. This convergence criterion is explained in the feature set evolution subsection of Section 4, where experimental results are also reported.

IV. Experimental study

In this section we report the experimental setup and results of our feature selection method proposed in Section 3. This method is applied to several microarray data sets described in Section 4.1. Four evaluation metrics, namely the classification performance, the stability of the selected genes and the algortithm convergence test are defined respectively.

A. Datasets

Three high dimensional data sets are used in our experimental study. The task in the Bladder cancer dataset is the clinical classification of bladder tumors using microarrays [13]. A list of 11 a priori relevant features, markers, are collected from the literature, looking systematically at the Pubmed literature on markers of recurrence and progression of bladder cancer.

The classification task in DLBCL, standing for diffuse large Bcells, is the prediction of the tissue types [14], where two genes previously known as clinical markers are used to discriminate DLBCL tissues from Follicular Lymphomas: Transferrin Receptor (TR) and Lactate Dehydrogenase A (LDHA).

We analyzed the microarray dataset, malignant pleural mesothelioma and lung adenocarcinoma gene expression database [15]. This Affymetrix Human GeneAtlas U95Av2 microarray dataset aims to test expression ratio-based analysis to differentiating between MPM and lung cancer. [16] collected prior knowledge from any proven information about lung adenocarcinoma related genes in the literature. Eight significant genes are considered (CXCLI, IL-18, AKAP12, KLF6, AXL ,MMP-12 ,PKP3 and CYP2A13).

Table I summarizes the characteristics of the three datasets, namely the number of samples, the initial dimension of the input space and the number of a priori relevant features.

B. Classification performance and stability

We use 10-fold stratified cross-validation to predict the classification performance on three data sets with the selected feature sets, obtained in 10 iterations. Classification accuracy is defined as the proportion of correct results that a classifier achieved. This metric is important and always used to evaluate feature selection algorithms for classification tasks. However, it is not sufficient given that there is no best way to evaluate any system, but different metrics give us different insights into how a feature selection algorithm performs.

The second important evaluation criterion used in our study is stability. Stability is defined as the sensitivity of a method to variations in the training set. The stability of a feature selection algorithm is the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution [2]. Stability quantifies how different training sets affect the feature preferences. [17] cites three sources that may cause instability of feature selection in biomarker discovery. Instability occurs because classic feature selection methods often ignore stability in the algorithm's design. The existence of multiple sets of true markers and the small number of samples in high-dimensional data are two other sources of feature selection instability. The motivation for investigating the stability of feature selection algorithms came from the need to provide application domain experts with quantified evidence that the selected features are relatively robust to variations in the training data. This need is particularly crucial in biological applications, e.g. genomics, DNA-micorarrays, proteomics and mass spectrometry. These applications are typically characterized by high dimensionality. The goal is to output a small set of highly discriminatory features on which biomedical experts will subsequently invest considerable time and research effort. Measuring stability requires a similarity measure for feature preferences that will measure to which extent K sets S of s selected features share common features. Those sets can typically be produced by selecting features from different samples of the data. [18] proposed the following stability index

$$Stab(S_1,..,S_K) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} (|S_i \cap S_j| - \frac{s^2}{d})/(S - \frac{s^2}{d}), \quad (1)$$

where d is the total number of features, and S_i, S_j are two feature sets built from different partitions of the training samples. This index satisfies $-1 < Stab \leq 1$ and the greater is its value the larger is the number of commonly selected features in various sets. A negative stability index means that feature sets sharing common features are mostly due to chance.

C. Feature set evolution

A crucial point to consider in evaluating our proposed feature selection algorithm is do we have a monotonic increase in the selected features (the R_n vector)? i.e. as we move from step n to $n+1$ do we always have $R_n \subseteq R_{n+1}$? This obviously depends on the behavior of the semi-supervised learning and the feature selection algorithm that we have selected for stages one and two. The convergence behavior of the two-step algorithm is measured by the quantity : $\frac{|R_n \cap R_{n+1}|}{|R_n \cup R_{n+1}|}$ as a function of n. This quantity is equal to zero when there are no common features between iteration $n+1$ and n and to 1 when there is no difference between the feature sets selected respectively in iteration $n+1$ and iteration n, meaning that the algorithm has converged. Another advantage of studying the algorithm's convergence is to use it as a stopping criterion for the feature selection process. The convergence scores show that for the three data sets, SS-L2AROM feature selection algorithm converges since the selected feature set becomes stable after a maximum of 6 iterations for Bladder cancer data set, 10 iterations for DLBCL data set and a maximum of 8 iterations for Lung cancer data set.

D. *Experimental results*

The proposed algorithm, SS-L2AROM, is compared with PS-l2-AROM, which as described before considers prior knowledge, and with SVM.RFE, which is a wrapper approach that does not integrate prior knowledge in the feature selection process. Table II presents the results of applying of the three algorithms on the three data sets, where N denotes the number of selected features.

Classification and stability results

Bladder cancer			
N	SS-L2AROM	PS-L2AROM	SVM-RFE
10	83.87- 0.8908	80.65- 0.7458	80.65- 0.7280
20	87.10- 0.8099	87.10- 0.6645	90.32- 0.7539
30	87.10- 0.8107	87.10- 0.6641	93.55- 0.7569
40	90.32- 0.8125	90.32- 0.6403	90.32- 0.7438
50	93.55- 0.8251	93.55- 0.6701	90.32- 0.7510
60	96.77- 0.8264	93.55- 0.6786	93.55- 0.7575
70	96.77- 0.8297	93.55- 0.6900	87.10- 0.7547
80	96.77- 0.8305	93.55- 0.6976	93.55- 0.7452
90	96.77- 0.8343	93.55- 0.7170	90.32- 0.7481
100	96.77- 0.8454	93.55- 0.7229	90.32- 0.7523
DLBCL			
N	SS-L2AROM	PS-L2AROM	SVM-RFE
10	92.21 - 0.8188	93.51 - 0.6996	67.53 - 0.4802
20	92.21 - 0.8448	94.81 - 0.7756	83.12 - 0.4771
30	94.81 - 0.8581	90.91 - 0.8215	88.31 - 0.4756
40	90.91 - 0.8863	92.21 - 0.8348	79.22 - 0.4744
50	93.51 - 0.8868	93.51 - 0.8753	81.82 - 0.4734
60	94.81 - 0.8855	90.91 - 0.8695	84.42 - 0.4677
70	96.10 - 0.8825	93.51 - 0.8722	87.01 - 0.4717
80	94.81 - 0.8868	94.81 - 0.8796	88.31 - 0.4708
90	94.81 - 0.8800	94.81 - 0.8853	85.71 - 0.4700
100	93.51 - 0.8880	94.81 - 0.8851	88.31 - 0.4663
Lung cancer			
N	SS-L2AROM	PS-L2AROM	SVM-RFE
10	98.34 - 0.8065	98.90 - 0.6508	91.16 - 0.7109
20	99.45 - 0.7941	99.45 - 0.7329	93.92 - 0.7552
30	99.45 - 0.8396	99.45 - 0.8099	96.13 - 0.7282
40	100 - 0.8250	99.45 - 0.7826	95.03 - 0.7191
50	100 - 0.8385	100 - 0.7680	93.37 - 0.7242
60	100 - 0.8605	100 - 0.7825	94.48 - 0.7302
70	100 - 0.8602	100 - 0.7874	95.58 - 0.7353
80	100 - 0.8541	100 - 0.7853	93.92 - 0.7408
90	100 - 0.8565	100 - 0.7789	96.13 - 0.7391
100	100 - 0.8640	100 - 0.7912	97.24 - 0.7377

TABLE II: Classification performance coupled with feature selection stability on Bladder cancer, DLBCL and Lung cancer data sets.

For Bladder cancer data set, SS-l2AROM gives the best classification performance with the best result obtained with a subset of 60 selected features (96.77%). SS-l2AROM gives also excellent stability results compared to PS-L2AROM and SVM-RFE and this is visibly clear in Fig. 1. The best stability value is 0.8908 and obtained with 10 features.

For DLBCL and Lung cancer data sets, SS-l2AROM still yields the best classification results as noticed in Table II. PS-L2AROM is a competitor algorithm concerning classification results both on DLBCL and Lung cancer data sets, but not on stability results where the outperformance of SS-l2AROM is clearly visible specially for Lung cancer data set (see Fig. 3). The stability behaviour of SVM-RFE is modest for all the experimented data sets. Thus, in most cases prior knowledge improves classification performance and stability results.

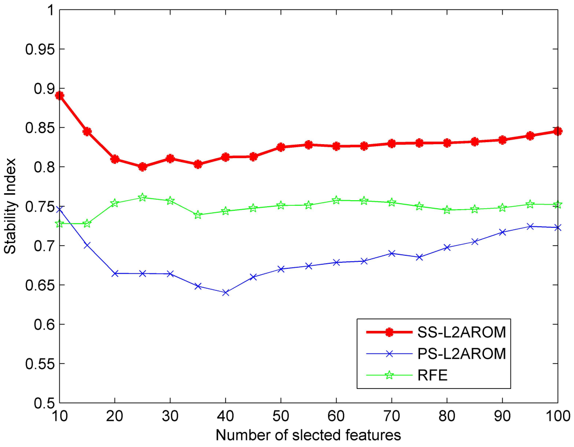

Fig. 1: Feature selection stability with Kuncheva Index on Bladder cancer data set

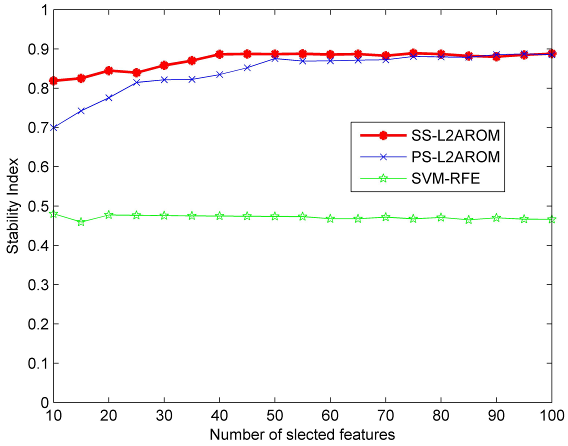

Fig. 2: Feature selection stability with Kuncheva Index on DLBCL data set

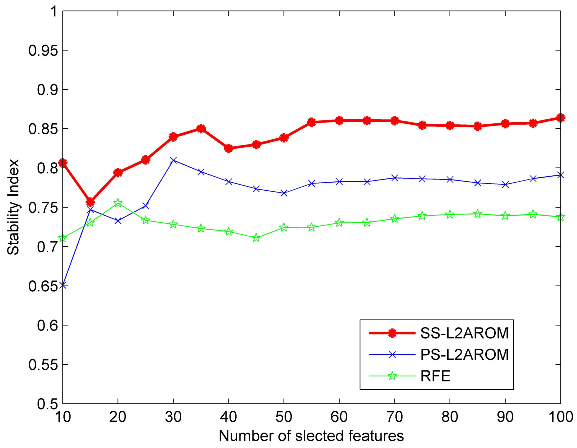

Fig. 3: Feature selection stability with Kuncheva Index on Lung cancer data set

From this empirical study, we deduce that algorithms which incorporate prior knowledge have a better classification accuracy than the other feature selection algorithms. This is not always the case for the stability of feature selection, but our proposed method, namely SS-l2AROM, is also advantageous in this respect. Consequently, considering background knowledge about features is very important and beneficial to guide the feature selection process. Moreover, taking advantage of this prior knowledge to extend the set of a priori relevant features in a pre-processing phase of feature selection further improves both classification and feature selection stability.

V. Conclusion

We propose a robust feature selection method, SS-L2AROM, based on semi supervised prior relevance learning. Prior knowledge about some dimensions known to be more relevant is incorporated as a means of guiding the feature selection process. The objective is to make use of a partial supervision on features assumed a priori to be more relevant, in order to select a robust feature set in an interactive manner. Iteratively we make use of the initial prior knowledge and the previously selected features to learn new relevant features by a semi supervised approach. The extended subset of relevant features is used as prior knowledge to be integrated in a second step to guide the feature selection process until an optimal number of features is obtained. Our proposed approach shows encouraging results both for improving the classification accuracy and for dealing with the instability problem in feature selection for high dimensional data. Experiments on three microarray data sets show that the partial supervision in SS-L2AROM improves both classification and stability performances compared to PS-L2AROM and SVM-RFE. Our proposed approach fits with any feature selection algorithm that can integrate prior knowledge.

VI. Acknowledgment

The registration fee for this conference has been sponsored by the Tunisian Decision Aid Society (TDAS): www.tdasociety.org.

References

[1] Y. Sun, S. Todorovic, and S. Goodison, "Local learning based feature selection for high dimensional data analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, pp. 1610–1626, 2010.

[2] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, 2007.

[3] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[4] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[6] T. Helleputte and P. Dupont, "Feature selection by transfer learning with linear regularized models," in *Proceedings of the of the European Conference on Machine Learning and Knowledge Discovery in Databases*. Berlin: Springer, 2009, pp. 533–547.

[7] E. Krupka and N. Tishby, "Incorporating prior knowledge on features into learning," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07)*, M. Meila and X. Shen, Eds., vol. 2. Journal of Machine Learning Research - Proceedings Track, 2007, pp. 227–234. [Online]. Available: http://jmlr.csail.mit.edu/proceedings/papers/v2/krupka07a/krupka07a.pdf

[8] A. Ben Brahim and M. Limam, "New prior knowledge based extensions for stable feature selection," in *Proceedings of the Sixth International Conference of Soft Computing and Pattern Recognition*. IEEE, 2014, pp. 306 –311.

[9] B. Taskar, M. F. Wong, and D. Koller, "Learning on the test data: Leveraging unseen features," in *Proc. ICML*, 2003.

[10] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, "Learning a meta-level prior for feature relevance from multiple related tasks," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 489–496. [Online]. Available: http://doi.acm.org/10.1145/1273496.1273558

[11] T. Helleputte and P. Dupont, "Partially supervised feature selection with regularized linear models," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 409–416.

[12] J. Weston, A. Elisseeff, B. Schlkopf, and P. Kaelbling, "Use of the zero-norm with linear models and kernel methods," *Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[13] L. Dyrskjot, T. Thykjaer, M. Kruhoffer, J. L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. F. Orntoft, "Identifying distinct classes of bladder carcinoma using microarrays," *Nat Genetics*, vol. 33, pp. 90–96, 2003.

[14] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, and D. S. Neuberg, "Diffuse large b(cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 9, pp. 68–74, 2002.

[15] G. Gordon, R. Jensen, L. Hsiao, S. Gullans, J. Blumenstock, S. Ramaswamy, W. Richards, D. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, pp. 4963–4967, 2002.

[16] P. Guan, D. Huang, M. He, and B. Zhou, "Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method," *Journal of Experimental and Clinical Cancer Research*, vol. 28, p. 103, 2009.

[17] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Computational Biology and Chemistry*, vol. 34, pp. 215–225, 2010.

[18] L. Kuncheva, "A stability index for feature selection," in *Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*, Innsbruck, Austria, 2007, pp. 390–395.