Kernel methods for mixed feature selection

Jérôme Paul and Pierre Dupont

Université catholique de Louvain - ICTEAM/Machine Learning Group
Place Sainte Barbe 2, 1348 Louvain-la-Neuve - Belgium

Abstract. This paper introduces two feature selection methods to deal with heterogeneous data that include continuous and categorical variables. We propose to plug a dedicated kernel that handles both kind of variables into a Recursive Feature Elimination procedure using either a non-linear SVM or Multiple Kernel Learning. These methods are shown to offer significantly better predictive results than state-of-the-art alternatives on a variety of high-dimensional classification tasks.

1 Introduction

Feature selection allows domain experts to interpret a decision model by reducing the number of variables to analyze. In medical studies for example, very high dimensional feature sets (e.g. gene expression data) are typically considered along with a few clinical features. These features can be continuous (e.g. blood pressure) or categorical (e.g. sex, smoker vs non-smoker).

To highlight important variables, a naive approach would transform heterogeneous data into either fully continuous or categorical variables before applying any standard feature selection algorithm. To get a continuous dataset, categorical variables can be encoded as numerical values. The specific choice of such numerical values is however arbitrary, introduces an artificial order between the feature values and can lead to largely different distance measures between instances [1]. Alternatively, continuous features can be discretized at the price of making the selection highly sensitive to the specific discretization [1].

To alleviate those limitations and to keep the heterogeneous (or mixed) nature of the data, we propose to use a dedicated kernel [2], designed to handle both kind of features. We will then perform feature selection according to the Recursive Feature Elimination (RFE) [3] mechanism either with non-linear SVMs or Multiple Kernel Learning (MKL) [4].

2 Related work

To the best of our knowledge, few selection techniques are specifically designed to deal with both continuous and categorical variables. An apparently natural approach would consider tree ensemble methods such as Random Forests (RF), since trees can be grown from both types of variables and these methods perform an embedded selection. RF were however shown to bias the selection towards variables with many values [5]. The cForest method has been introduced to correct this bias [5] but its computational complexity prevents its use when dealing with thousands of features.

Proc. of the *22th European Symposium on Artificial Neural Networks (ESANN'14)*, Bruges, Belgium, pp. 301-306, April, 2014.

An alternative method performs a greedy forward selection aggregating separate rankings for each type of variables into a global ranking [1]. The authors report improved results over those of the method proposed in [6], which is based on neighborhood relationships between heterogeneous samples.

Out of a total of p variables, categorical and continuous features are first ranked independently. Mutual information (MI) was originally proposed for those rankings but a reliable estimate of MI is difficult to obtain whenever fewer samples than dimensions are available. Instead we use the p-values of a t-test to rank continuous features and of a Fisher exact test for categorical ones. The two feature rankings are then combined into a global ranking by iteratively adding the first categorical or continuous variable that maximizes the predictive performance of a Naive Bayes or a 5-NN classifier (consistently with the choices made in [1]). The NN classifier uses the Heterogeneous Euclidian-Overlap Metric [7] between pairs of instances as follows:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{f=1}^{p} d_f(x_{if}, x_{jf})^2} \tag{1}$$

$$d_f(a,b) = \begin{cases} I(a \neq b) & \text{if } f \text{ is categorical} \\ \frac{|a-b|}{max_f - min_f} & \text{if } f \text{ is continuous} \end{cases} \tag{2}$$

where I is the indicator function. Unlike a simple 0/1 encoding of categorical features, this distance does not introduce an arbitrary order between feature values and preserves the original number of dimensions. We refer to these approaches as HFS^{NB} and HFS^{5NN} in the sequel.

The so-called clinical kernel proposed in [2] was shown to outperform a linear kernel for classifying heterogeneous data. It averages univariate subkernels defined for each feature and is closely related to the metric defined in equation (2):

$$k(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{p} \sum_{f=1}^{p} k_f(x_{if}, x_{jf}) \tag{3}$$

$$k_f(a,b) = 1 - d_f(a,b) \tag{4}$$

This kernel assumes the same importance to each original variable and corresponds to concatenating those variables in the kernel induced feature space. We show here the benefit of adapting this kernel for heterogeneous feature selection.

3 Methods

Section 3.1 briefly recalls the Recursive Feature Elimination (RFE) procedure. Section 3.2 details how to obtain a feature ranking from a non-linear SVM. Finally, section 3.3 sketches Multiple Kernel Learning, which offers an alternative way to rank variables with the clinical kernel.

3.1 Recursive feature elimination

RFE [3] is an embedded backward elimination strategy that iteratively builds a feature ranking by removing the least important features in a classification

model at each step. Following [8], one typically drops a fixed proportion (e.g. 20 %) of features at each iteration. RFE is most commonly used in combination with a linear SVM from which feature weights are extracted. However, it can be used with any classification model from which individual feature importance can be deduced.

Algorithm 1: Recursive Feature Elimination

```
R ← empty ranking
F ← set of all features
while F is not empty do
    train a classifier m using F
    extract variable importances from m
    remove the least important features from F
    put those features on top of R
return R
```

3.2 Feature importance from non-linear Support Vector Machines

In order to extract variable importance from a non-linear SVM, one can look at the influence on the margin of removing a particular feature. The margin is inversely proportional [9] to

$$W^2(\alpha) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{w}||^2 \tag{5}$$

where α_i and α_j are the dual variables of a SVM, y_i and y_j the labels of $\mathbf{x}_i$ and $\mathbf{x}_j$, out of n training examples, and k a kernel. Therefore, the importance of a particular feature f can be approximated [9] without re-estimating α by the following formula:

$$J_{SVM}(f) = |W^2(\alpha) - W^2_{(-f)}(\alpha)| \tag{6}$$

$$W^2_{(-f)}(\alpha) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i^{-f}, \mathbf{x}_j^{-f}) \tag{7}$$

where $\mathbf{x}_i^{-f}$ is the i^{th} training example without considering the feature f.

In this work, we propose to combine this feature importance with the RFE mechanism in order to provide a full ranking of the features. This method will be referred to as RFE^{SVM}.

3.3 Feature importance from Multiple Kernel Learning

MKL [4] learns an appropriate linear combination of K base kernels

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{K} \mu_m k_m(\mathbf{x}_i, \mathbf{x}_j) \qquad \text{s.t. } \mu_i \geq 0 \tag{8}$$

In this work, we adapt the clinical kernel (see equation (4)) with MKL to learn a non-uniform combination of the base kernels, each one associated to a single

feature. Because each base kernel has the same scale whether it is built on a continuous or categorical variable, μ_f can be seen as the importance $J_{MKL}(f)$ of feature f. The combination of RFE with this feature importance extracted from MKL will be referred to as RFE^{MKL}. It specifically uses the kernel weights μ_m (see equation (8)) to guide the search towards relevant features.

4 Experimental setting and datasets

In order to assess feature selection, we report predictive performances of classifiers built on selected variables as well as the stability of those feature sets. We make use of a resampling strategy consisting of 200 random splits of the data into training (90%) and test (10%). For each data partition, the training set is used to rank features and build predictive models using different numbers of features. The ranking is recorded and predictive performances are measured while classifying the test set. Average predictive performances are reported over all test folds and the stability of various signature sizes is computed from the 200 feature rankings.

Predictive performances are reported here in terms of balanced classification rate (BCR), which is the average between sensitivity and specificity. These metrics are particularly popular in the medical domain and BCR, unlike AUC, easily generalizes to multi-class with unbalanced priors. Selection stability is assessed here through the Kuncheva's index (KI) [10].

We report results on 5 binary classification datasets briefly described in the following table:

Name	Task	Continuous features	Categorical features	Class priors
Arrhythmia [11]	presence of cardiac arrhythmia	198	64	245/185
Heart [11]	presence of heart disease	6	7	164/139
Hepatitis [11]	survival to hepatitis	6	13	32/123
Rheumagene [12]	early diagnosis of arthritis	100	3	28/21
van't Veer [13]	breast cancer prognosis	4348	7	44/33

5 Experimental results

We report predictive performances obtained with a non-linear SVM using the clinical kernel reduced to the selected features from the various selection techniques considered. Similar results are obtained using RF, Naive Bayes or 5-NN as final classifier.

Results presented in Figure 1 show that the proposed selection methods RFE^{SVM} and RFE^{MKL} generally outperform the greedy forward selection

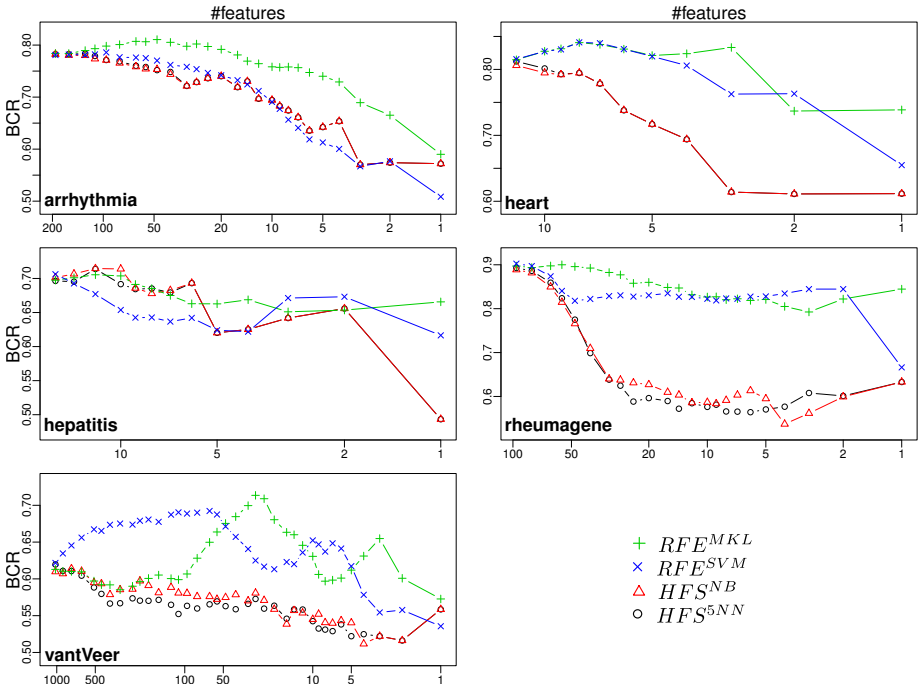

Fig. 1: Predictive performances with respect to the number of selected features.

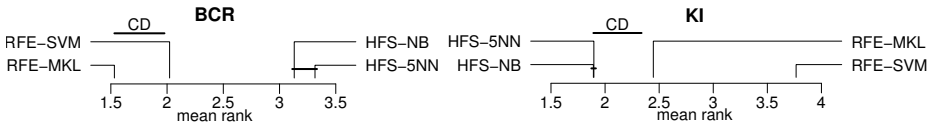

Fig. 2: Nemenyi critical difference diagrams : comparison of the predictive performances and stability of the four algorithms.

approach described in Section 2. The BCR improvement over the two HFS approaches tends to be larger for high dimensional datasets. The BCR differences are highly significant according to a Friedman test [14] across all datasets and all feature set sizes (p-val $< 10^{-30}$). A Nemenyi post-hoc test shows that RFE^{MKL} performs best, followed by RFE^{SVM}, and that both approaches significantly outperform the HFS methods.

Nemenyi critical difference (CD) diagrams [14] are reported in Figure 2. The highly significant gains in predictive performance (BCR) obtained with the proposed methods come at the price of significantly lower stability results (KI). This is most likely due to the multivariate nature of the selection used as opposed to the HFS methods aggregating univariate rankings known to be stable but potentially less predictive.

6 Conclusion and future work

We introduce two heterogeneous feature selection techniques that combine Recursive Feature Elimination with variable importances extracted from a non-linear SVM or through MKL. These methods use a dedicated kernel combining continuous and categorical variables. Experiments show that they improve predictive performances of state of the art methods, especially for high dimensional datasets. Improving the selection stability of those methods is part of our future work, possibly by resorting to an ensemble procedure [8].

References

[1] G. Doquire and M. Verleysen. An hybrid approach to feature selection for mixed categorical and continuous data. In *KDIR*, pages 394–401, 2011.

[2] A. Daemen and B. De Moor. Development of a kernel function for clinical data. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 5913–5917, 2009.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[4] G. Lanckriet, T. De Bie, N. Cristianini, M. Jordan, and W. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[5] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

[6] Q. Hu, J. Liu, and D. Yu. Mixed feature selection based on granulation and approximation. *Knowledge-Based Systems*, 21(4):294 – 304, 2008.

[7] R. Wilson and T. Martinez. Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

[8] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2010.

[9] I. Guyon. *Feature extraction: foundations and applications*, volume 207. Springer, 2006.

[10] L. Kuncheva. A stability index for feature selection. In *AIAP'07: Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference*, pages 390–395, Anaheim, CA, USA, 2007. ACTA Press.

[11] A. Frank and A. Asuncion. UCI ML repository - http://archive.ics.uci.edu/ml, 2010.

[12] I. Focant, D. Hernandez-Lobato, J. Ducreux, P. Durez, A. Toukap, D. Elewaut, F. Houssiau, P. Dupont, and B. Lauwerys. Feasibility of a molecular diagnosis of arthritis based on the identification of specific transcriptomic profiles in knee synovial biopsies. *Arthritis and Rheumatism*, 63(10, suppl.):abstract 1927:S751, 2011.

[13] L. van 't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.

[14] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006.

Acknowledgements

Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI) funded by the Fonds de la Recherche Scientifique de Belgique (FRS-FNRS).