A simple and efficient algorithm for variable ranking and redundancy detection

Oscar Luaces, Juan José del Coz
Machine Learning Group
Artificial Intelligence Center
University of Oviedo at Gijón

Universidad
de Oviedo

Artificial
Intelligence
Center

Outline

- Introduction
- State-of-the-art in variable ranking:
  - wrappers
  - filters
- Our approach: SPE-ranker
  - Simplified Polynomial Expansion + correlation-based ranking
  - Redundancy detection
- Experimental results

Introduction

- Variable selection plays a very important role in most (if not all) learning tasks to improve the accuracy and readability of learned models
- A usual approach to linearize the search for an optimal subset of input variables is to previously construct a ranking
- Most ranker variables fall in one of these categories:
  - wrappers
  - filters
- In this work we present a filter called SPE-ranker:
  - Simplified Polynomial Expansion to deal with nonlinear problems
  - Correlation-based ranking criterion
  - Gram-Schmidt orthogonalization to detect redundancy

Some state-of-the-art rankers: SVM wrappers

- [Rakotomamonjy, 2003]: orders the list of variables according to the influence of the variations of their weights

$$R_{\mathtt{RM}}(i) = |\nabla_i \|\boldsymbol{w}\|^2| = \left| \sum_{k,j} \alpha_k \alpha_j y_k y_j \frac{\partial K(\boldsymbol{s} \cdot \boldsymbol{x}_{k,\cdot}, \boldsymbol{s} \cdot \boldsymbol{x}_{j,\cdot})}{\partial s_i} \right|, \quad i = 1, \ldots, m$$

- [Degroeve et al., 2002]: variables are ordered with respect to the loss in predictive performance when they are removed

$$R_{\mathtt{DM}}(i) = \sum_k y_k \cdot \sum_j \alpha_j y_j K(\boldsymbol{x}^{(i)}_{j,\cdot}, \boldsymbol{x}^{(i)}_{k,\cdot}), \ i = 1, \ldots, m$$

- [Weston et al., 2001] (R^2W^2): expressing the kernel as $K_{\boldsymbol{\sigma}}(\boldsymbol{x}, \boldsymbol{x'}) = K(\boldsymbol{x} * \boldsymbol{\sigma}, \boldsymbol{x'} * \boldsymbol{\sigma})$ the authors propose to find σ by minimizing the expected error bound using gradient descent

$$EP_{err} \leq \frac{1}{n} E\left\{ \frac{R^2}{M^2} \right\} = \frac{1}{2} E\{R^2 W^2(\alpha^0)\}$$

where R is the radius of a sphere containing the examples and M is the margin

Some state-of-the-art rankers: filters

- [Stoppiglia et al., 2003]:
  - Select the variable that best explains the target output

$$R_{\texttt{SM}}(i) = \cos^2(\boldsymbol{x}_i, \boldsymbol{y}) = \frac{\langle \boldsymbol{x}_i, \boldsymbol{y} \rangle^2}{\|\boldsymbol{x}_i\|^2 \|\boldsymbol{y}\|^2}, \quad i = 1, \ldots, m$$

  - Using the Gram-Schmidt orthogonalization, project the column vectors of the remaining input variables and the target output onto the space spanned by the column vectors previously selected
  - Orthogonalization is used to discard the part of the target concept already explained by previously selected variables, so the variable selected in the next iteration will be the most relevant with respect to what is not yet explained

Some state-of-the-art rankers: filters

- [Guyon et al., 2003]: similar to Stoppiglia's method but replacing the ranking criterion by a slightly modified version of the classical *Relief* [Kira and Rendell, 1992]
- [Yu and Liu, 2004] (*FCBF*): relevance analysis + redundancy detection
  - Relevance score: computed as the *symmetric uncertainty* with respect to the target output

$$SU(\boldsymbol{x}_i, \boldsymbol{x}_j) = 2\left[\frac{IG(\boldsymbol{x}_i|\boldsymbol{x}_j)}{H(\boldsymbol{x}_i) + H(\boldsymbol{x}_j)}\right]$$

  - Redundancy detection: iterative process that removes those variables for which x_i forms an *approximate Markov blanket*:
    - a variable x_i forms an approximate Markov blanket for any other variable x_j if and only if $SU(x_j,y) \geq SU(x_i,y)$ and $SU(x_i,x_j) \geq SU(x_i,y)$

SPE-ranker

- We propose a *simplified polynomial expansion* (SPE) to transform the input space *X* into a new feature space φ(*X*), which hopefully will have a quasi-linear relationship with the class

$$\varphi(x_1, \ldots, x_m) = (x_1, \ldots, x_m, x_1^2, \ldots, x_m^2, \ldots, x_1^d, \ldots, x_m^d)$$

- Correlation-based ranking criterion: the ranking score of a variable x_i will be given by the best score among those obtained by the powers of x_i

$$R_{\text{SPE}}(i) = \max_{j=1,\ldots,d} \rho^2(\boldsymbol{x}_i^j, \boldsymbol{y}), \quad i = 1, \ldots, m$$

- Standardizing and normalizing the feature vectors the ranking criterion can be expressed as a scalar product

$$\begin{aligned} R_{\text{SPE}}(i) &= \max_{j=1,\ldots,d} (\rho^2(\boldsymbol{x}_i^j, \boldsymbol{y})) = \max_{j=1,\ldots,d} (\cos^2(\hat{\boldsymbol{x}}_i^j, \hat{\boldsymbol{y}})) = \\ &= \max_{j=1,\ldots,d} \left(\frac{\langle \hat{\boldsymbol{x}}_i^j, \hat{\boldsymbol{y}} \rangle^2}{\|\hat{\boldsymbol{x}}_i^j\|^2 \|\hat{\boldsymbol{y}}\|^2} \right) = \max_{j=1,\ldots,d} (\langle \hat{\boldsymbol{x}}_i^j, \hat{\boldsymbol{y}} \rangle^2), \quad i = 1, \ldots, m \end{aligned}$$

SPE-ranker

- Redundancy detection:

  1. We select the column vector x_i corresponding to the leading position of the ranking. Then a column matrix q is built with the standardized and then normalized version of x_i
  2. The rest of the input variables (column vectors) are standardized and orthogonally projected onto the space spanned by q
  3. The norms of the resulting projections are compared with the norms of the original standardized vectors. Those projected input variables whose norm decrease more than a given threshold δ are considered redundant and, therefore, they are removed
  4. The projection of the next input variable (following the ranking) is normalized and included in q. If a stopping criterion is not met, then we repeat this process starting at step 2

Experimental results

- We define a problem space where each data set is randomly generated according to the following six parameters:
  - Number of input variables (*m*)
  - Number of examples (*n*)
  - Number of relevant input variables (*r*)
  - Degree (*d*)
  - Redundancy (η)
  - Noise (σ)
- *Single experiment*: Hold-out with 30 artificially generated data sets for each point of the problem space
- Varying the parameters of the problem space we have made 2x160 single experiments (2x4800 random data sets) to analyze the effect of redundancy and the number of irrelevant variables

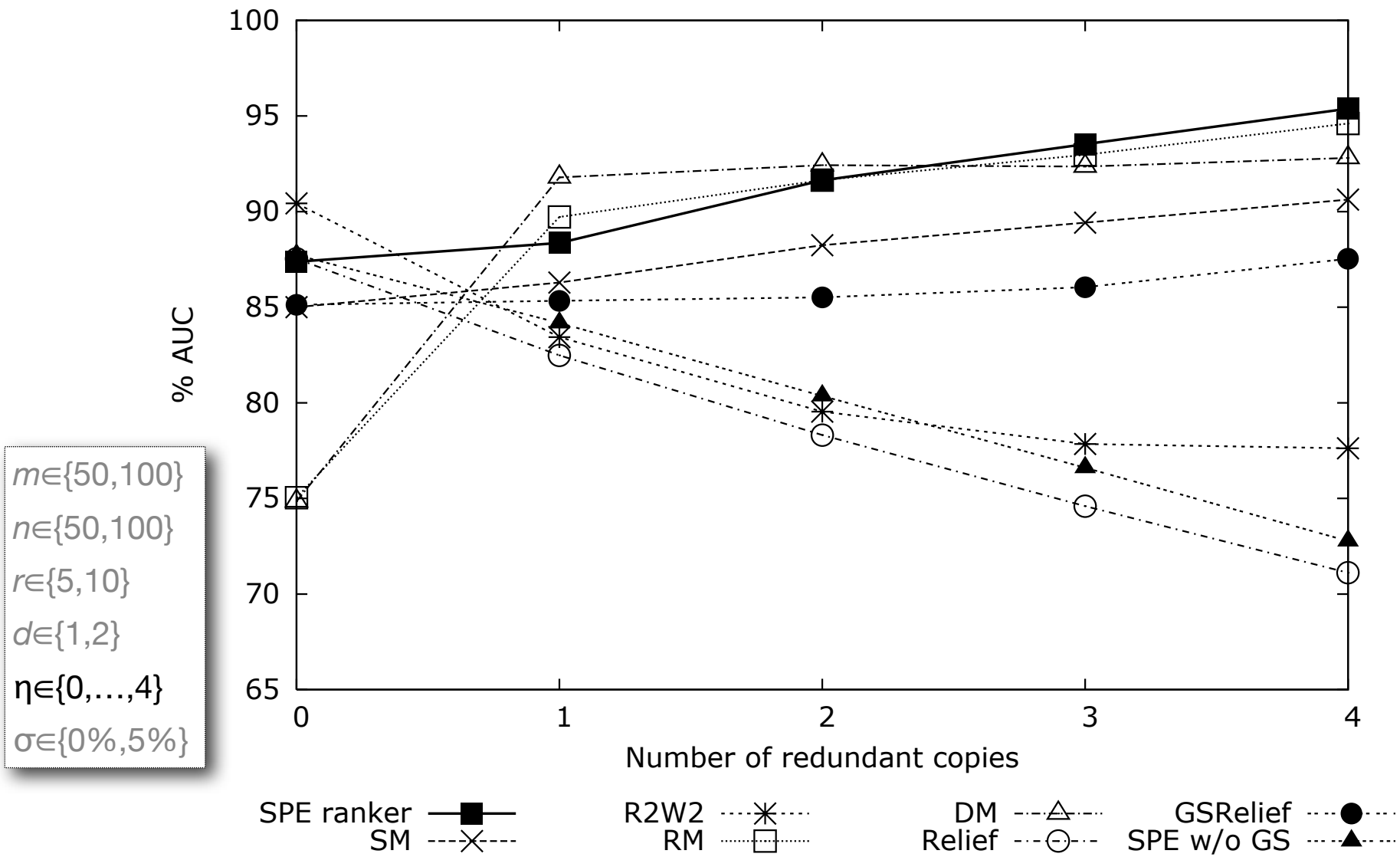
Experimental results: performance measure

- Given that we know the true relevancy of the input variables we can use a ROC-inspired setting [Jong et al., 2004] to evaluate the quality of the obtained rankings
  - We draw the ROC-FR curve with the points {(FPR(i),TPR(i)), i=1,…, m}, where TPR(i) (respectively FPR(i)) is calculated as the fraction of true (false) relevant variables whose position in the ranking is higher than i

    Since we use one-to-one redundancy, we only consider as relevant either an original relevant variable or one of its redundant copies (if there is more than one), whatever occurs first starting from the top of the ranking

  - We will use the area under the ROC-FR curve (AUC-FR) as an indicator of the quality of a ranking
- A running time comparison is also included

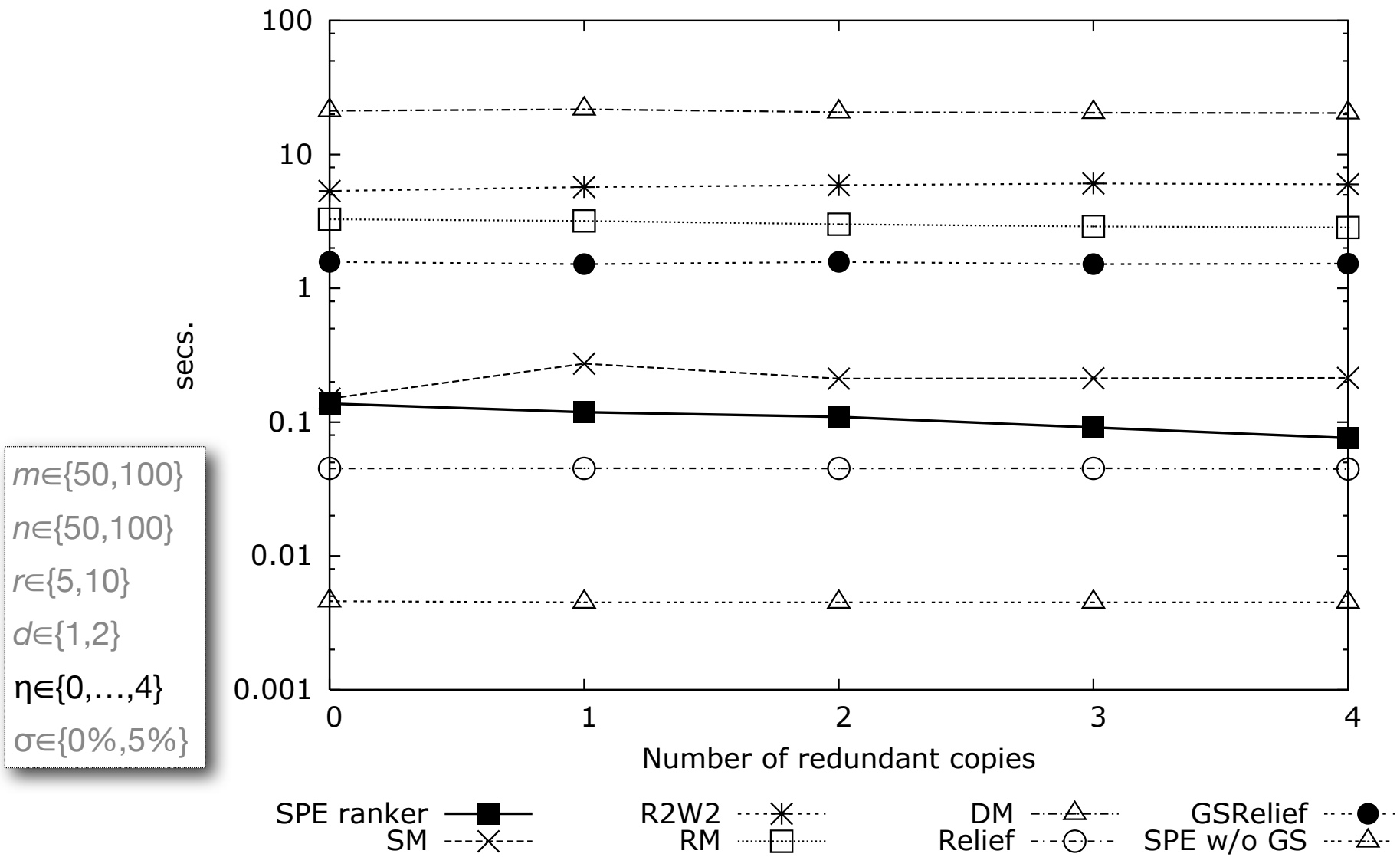
Experimental results
Varying redundancy: AUC-FR



Universidad de Oviedo

Artificial Intelligence Center

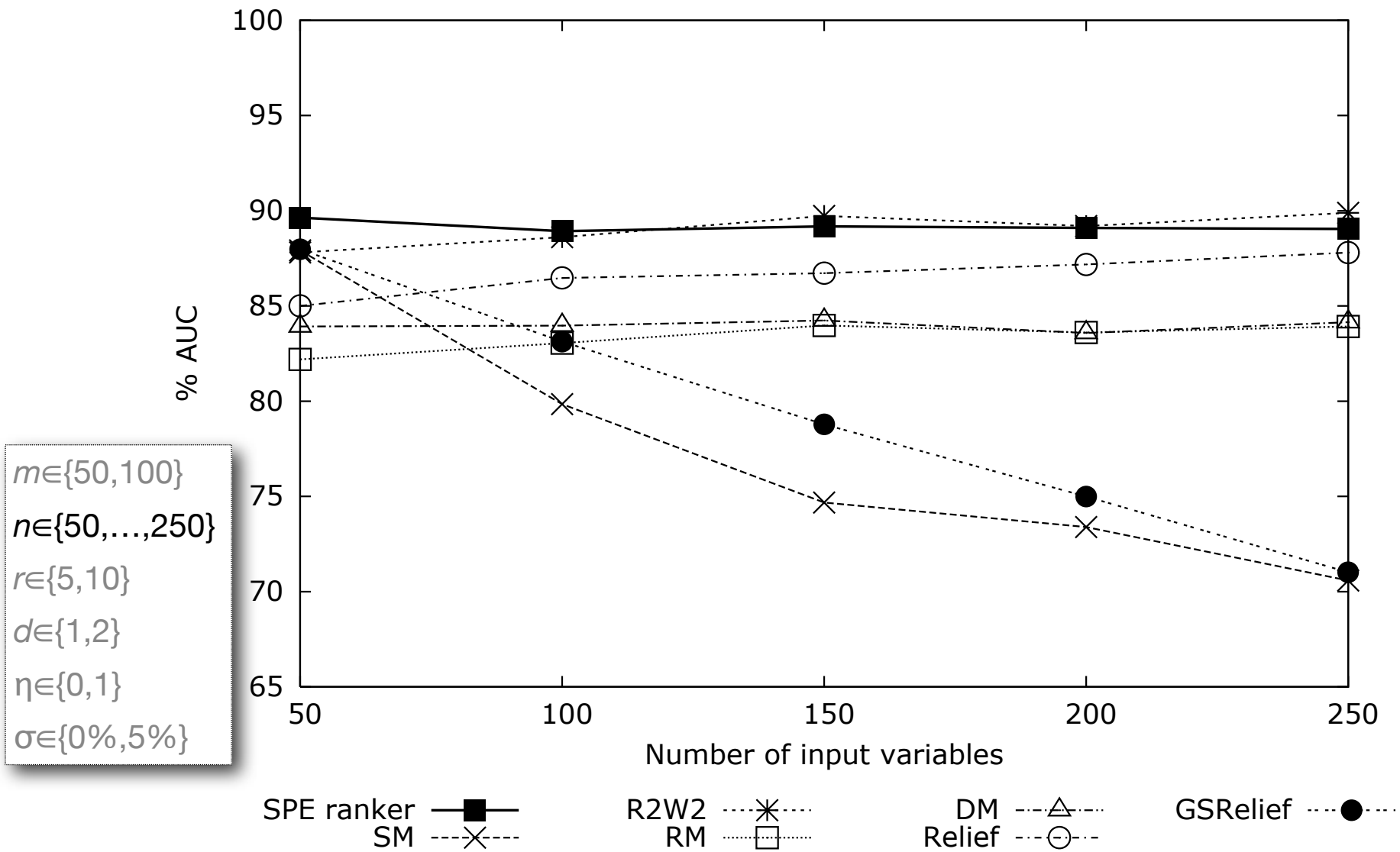
Experimental results
Varying redundancy: running time



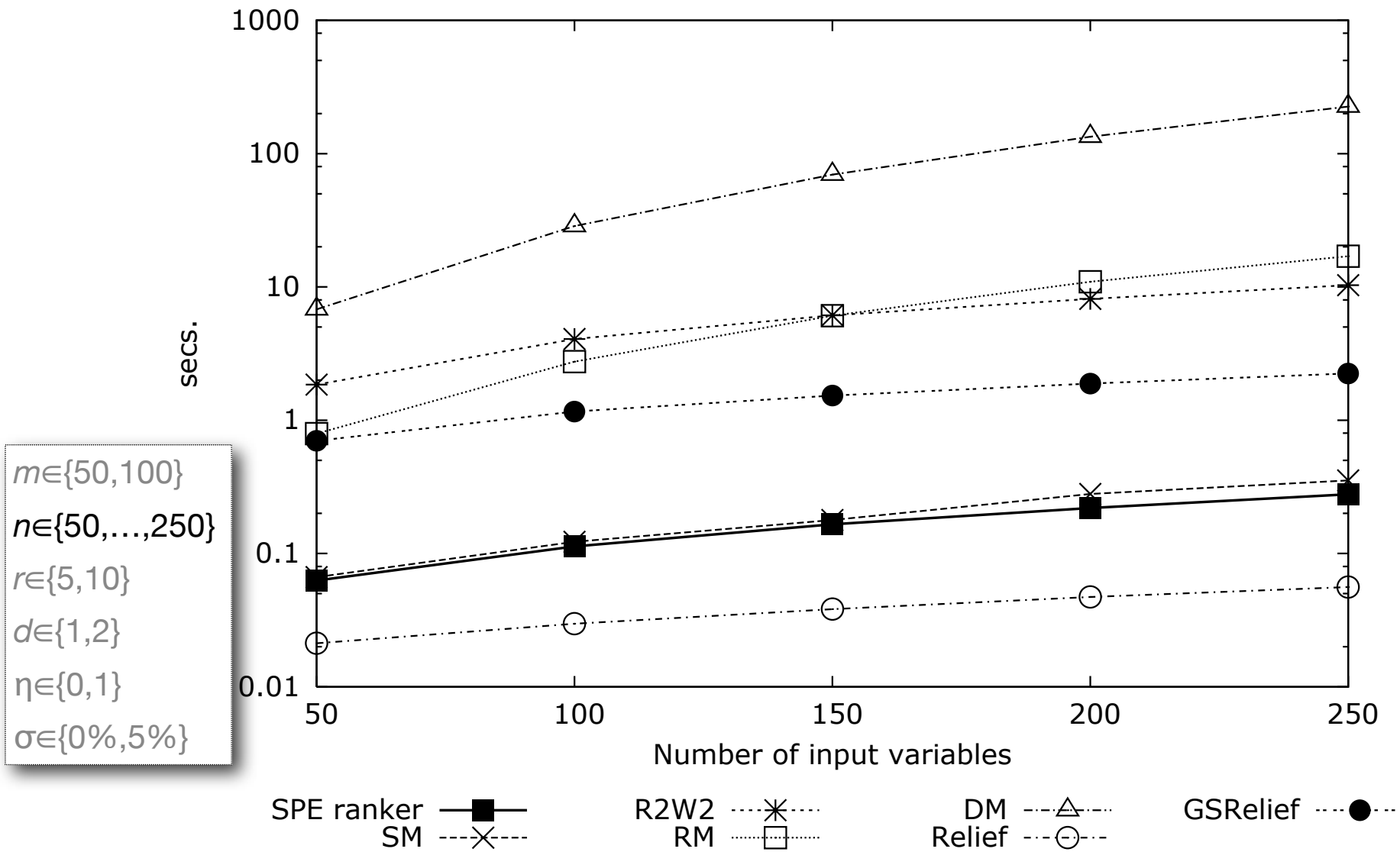
Universidad de Oviedo

Artificial Intelligence Center

Experimental results
Varying # of irrelevants: AUC-FR



Experimental results
Varying # of irrelevants: running time



Conclusions

- We have presented SPE-ranker, a simple approach for input variable ranking

Simplified Polynomial Expansion
+
Correlation
+
Orthogonalization

that deals explicitly with non linear problems and redundancy

- Good performance on artificially constructed data sets
  - in terms of AUC-FR
  - in running time