ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 286, 2013

Mariusz Kubus[*]

SOME REMARKS ON FEATURE RANKING BASED WRAPPERS

Abstract. One of the approaches to feature selection in discrimination or regression is learning models using various feature subsets and evaluating these subsets, basing on model quality criterion (so called wrappers). Heuristic or stochastic search techniques are applied for the choice of feature subsets. The most popular example is stepwise regression which applies hill-climbing. Alternative approach is that features are ranked according to some criterion and then nested models are learned and evaluated. The sophisticated tools of obtaining a feature rankings are tree based ensembles. In this paper we propose the competitive ranking which results in slightly lower classification error. In the empirical study metric and binary noisy variables will be considered. The comparison with a popular stepwise regression also will be given.

Key words: feature selection, wrappers, feature ranking.

I. INTRODUCTION

Data mining techniques are used as a knowledge discovery tools for huge datasets. Researcher often has no prior knowledge on the proper specification of the model nor on most informative features which influence examined phenomenon (which is represented by dependent variable in regression and discrimination). One can point several methods, i.e. tree based ensembles (Gatnar 2008), which classify very accurately future objects (out of training sample) but they work as a *black box*, not leaving much interpretation possibility. That is why the linear models are still attractive. The goal of the analysis is often not only to learn the accurate classifier but to discover the hidden relations in the data. The agent of the mobile phone network or the agent of the insurance company would like to know the reasons of the customer migrations. The investors would like to know the rules predicting the bancruptcy of the companies, and the doctors would like to make a diagnosis as soon as possible to choose the right way of the treatment. Using the contemporary clinical technologies it is sometimes done basing on thousands gene expressions obtained from DNA code.

[*] Ph.D., Department of Mathematics and Applied Computer Science, Opole University of Technology.

There are usually many noisy variables in real word domains and application of the discrimination methods without feature selection leads to overfitted and instable models. It is a well known fact that complex models which suite to the training data very hard do not classify the future objects very well (see i.e. Hastie *et al.* (2009), p. 38). On the other hand, too simple models do not extract all information from the data what also results in high test error. As the example consider the default rule of classification. The new example is classified to the group with the highest prior probability which is estimated using training sample. In that case no information carried by explanatory variables is used. Therefore, the key idea in predictive modelling is to choose the model of the right complexity, which is the compromise between these two extremes. The measure of the complexity in the linear model is usually the number of the parameters. It is equivalent to the number of variables if one does not include interaction terms or other functions of the original input variables to the model specification.

The methods of feature selection are currently classified into three groups: filters, wrappers and embedded methods (see i.e. Blum and Langley (1997); Guyon *et al.* (2006)). All of them perform a search in the space of all possible subsets of variables. The differences between them are whether the search is outside or inside the learning algorithm and whether the criterion is connected with a model or not. Wrappers perform a search outside the learning algorithm (unlike the embedded methods) and criterion is strictly connected with the model (in contrast to filters). Various wrapper approaches vary mainly in a search strategy. Feature ranking based wrappers – which we focus on in this paper – are considered to be less prone to overfitting (Ng 1998).

The goal of this paper is to propose feature ranking which is obtained from regularized linear regression model. From wrapper methodology point of view it will be compared to the rankings obtained from tree based ensembles. In the empirical study metric and binary noisy variables will be considered.

II. WRAPPERS

Suppose we are given a vector of input variables $\boldsymbol{X} = (X_1,...,X_p)$ and binary response Y, which categories will be called classes. Without the loss of generality we assume that $X_1,...,X_p$ are metric or binary. In the case of ordered or nominal variables with many categories, one can transform them to the set of dummy variables. The task of discrimination is to learn the model given the training set:

$$\{(\boldsymbol{x_1}, y_1),...,(\boldsymbol{x_N}, y_N) : \boldsymbol{x_i} \in \boldsymbol{X} = (X_1,...,X_p), y_i \in \{0,1\}, i \in \{1,...,N\}\} \quad (1)$$

and to use this model for classification of the new object $\boldsymbol{x}$ with unknown response y. We consider the situation when p is quite large and there are some noisy variables in the data which carry no significant information on the differences between classes. The presence of such variables in the data can decrease the accuracy of the model. The task of feature selection is to identify the subset $S_{\mathbf{X}} \subset \boldsymbol{X}$ so that the model learned with a use of $S_{\boldsymbol{X}}$ would be not worse than the model learned with a use of all input variables $\boldsymbol{X}$ (due to some model quality criterion).

Feature selection can be formulated as the search in the space of all possible feature subsets. As the exhaustive search (evaluating of all possible subsets) is computationally expensive and even not recommended from overfitting problem point of view, the heuristic or stochastic search is usually applied. For the brief survey see i.e. Reunanen (2006). The main idea of the search is to point which subsets of the input variables are worth to be evaluated. The second important issue is the computational cost. The search is controlled by the chosen criterion. In the wrapper methodology the feature subsets are evaluated by a model quality.

The most popular search strategy is hill-climbing (also known as greedy search). In every iteration, the current feature subset is modified so that the criterion would be improved. The simplest form of modification is to add or remove one variable from the current subset. In linear model literature such procedure is known as stepwise regression, but note that it is more general. It can be applied with a use of various learning algorithms and even with the criterion independent of the model (multivariate filters). The search can be performed in two directions. Starting from the empty subset of variables one can add one variable in every iteration – so called forward selection. One can also start from the full set of the input variables $\boldsymbol{X}$ and discard one variable in every iteration – so called backward elimination. The combination of both is also possible. In this way a variable can be added or removed in every iteration. Sometimes the term stepwise regression is reserved in the literature for that bi-directional procedure.

Commonly used model quality criterion is classification error estimated via cross-validation. The alternative is to use the information criteria what does not require splitting the data into training and test samples. The second approach considerably decreases the cost of computations. Usually the search is performed as long as the quality of the model is improved. It is natural stopping criterion which decides that forward selection is clearly faster in the case of high dimension of the feature space. That is because the learned models use smaller subsets of the variables.

The alternative approach to the mentioned search strategies is feature ranking. The main objective of the ranking is to constitute the search path in the wrapper methodology. Thus, the algorithm consists of two steps. Variables are ordered according to the chosen criterion and then nested models are learned, evaluated and the best one is chosen. The simplest proposition for creating of the ranking is using t-test statistic. The problem appears when there are nominal variables in the data. Then one can use the homogeneity measures which are applied in classification trees, i.e. information gain (the survey of these measures is given in (Gatnar 2001)). Nevertheless, still one deals with univariate approach. The sophisticated tools for creating the feature ranking are tree based ensembles, i.e. boosted trees (Freund and Schapire 1996) or random forests (Breiman 2001). These rankings represent multivariate approach because the values of the homogeneity measure for all variables are aggregated over all nodes in every tree. As the ensembles are very effective tools for classification as well as robust against noisy variables one can suspect that these rankings will reflect the importance of the variables very well. In the next section we propose ranking which will be obtained from regularized linear model.

III. REGULARIZED LINEAR REGRESSION

The parameters of the regularized linear regression model are estimated by minimizing the sum of loss function and penalty component:

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \left(L(y_i, b_0 + \sum_{j=1}^{p} b_j x_{ij}) + \lambda \cdot P(\boldsymbol{b}) \right). \tag{2}$$

The first component represents the goodness of fit and in the multiple regression it is usually square loss function. The second one supplies the possibility of controlling the complexity of the model. Various penalty formulas were proposed in the literature, i.e. in ridge regression (Hoerl and Kennard 1970) or LASSO (Tibshirani 1996) to mention the most popular. We focus on *elastic net* (Zou and Hastie 2005):

$$P_{\alpha}(\boldsymbol{b}) = \sum_{j=1}^{p} \left(\alpha b_j^2 + (1-\alpha)\left| b_j \right| \right), \tag{3}$$

which combines both. The penalty affects shrinking of the coefficients to zero. In practice, some of the coefficients are equal to zero what supplies the feature selection effect. The penalty parameter λ decides about the amount of shrinkage and it is usually tuned via cross-validation (see the experimental study in (Kubus 2011)).

The linear model with regularization was primary proposed for multiple regression but it can be easily adopted to discrimination in the case of two classes. Then the classes are coded by 0 and 1 which are interpreted as the probabilities that observed object is from the class coded by 1. The model is known as a linear probability model (LPM). It is not attractive discrimination method (estimated posterior probabilities can extend the $[0,1]$ interval) but when the regularization term is included to the LPM, it better discards noisy variables than regularized logistic regression (see Kubus (2013)). The absolute values of the coefficients can be treated as the importance measure of the variables in discrimination task (of course if the variables are standardized). To obtain more stable solution we propose to apply 10-fold cross-validation. Thus, we take the absolute value from the sum of 10 estimates of the coefficient as the importance measure of the variable.

IV. EXPERIMENT

Four datasets from UCI Repository of Machine Learning Databases (Frank and Asuncion 2010) were used in the experiment (Tab.1).

Table 1. Datasets used in the experiment

Dataset	# observations	# variables	# classes
breast cancer	569	30	2
Ionosphere	351	33	2
Pima	768	8	2
Sonar	208	60	2

Source: *UCI Repository of Machine Learning Databases.*

Additionally, noisy variables were included in the original datasets. We added them in four ways:

1) 10 noisy variables from Bernoulli distribution (with equal fractions of 0 and 1),

2) 10 noisy variables from $N(0,1)$ (some of them were collinear),

3) 10 noisy variables as in 1) and 10 noisy variables as in 2),

4) 20 noisy variables from Bernoulli distribution with fraction of 1 equal to 20%.

In this way we obtained 16 datasets which we denoted DATASET _#, where # is the scheme of generating of noisy variables.

The goal of the experiment was to compare three rankings. We were interested in getting the answers to two questions. How the rankings affect the detection of noisy variables? How they affect the classification error?

They were denoted as follows:

EN – ranking obtained from *elastic net* (Zou and Hastie 2005) with additional use of cross-validation as proposed in section III,

RF – ranking obtained from random forests (Breiman 2001),

BT – ranking obtained from boosted trees (Freund and Schapire 1996).

To obtain the family of nested models we used logistic regression and model selection criterion was BIC. The results are summarised in the Table 2. Proposed ranking better identified noisy variables than BT but worse than RF. Nevertheless, classification error was almost always slightly lower when EN was applied. It seems to be in contradiction to overfitting phenomenon. RF leads to a little bit less complex models which result in higher classification errors. Note, however, that we count as noisy variables only those artificially added according the schemes 1-4. In fact, there can be noisy variables in the original datasets which were discarded by EN and not discarded by RF. We observed that models learned with a use of EN usually contained fewer variables.

Table 2. Mean numbers of noisy variables introduced into the models and classification errors estimated via 10-fold cross-validation (standard errors in brackets)

Dataset	**# noisy variables**			**cv errors**		
	EN	**RF**	**BT**	**EN**	**RF**	**BT**
breast cancer 1	0 (0)	0 (0)	0 (0)	**2.8 (0.8)**	4.7 (1.2)	4.0 (0.8)
breast cancer 2	0 (0)	0 (0)	0 (0)	**2.6 (0.6)**	4.7 (0.6)	3.3 (1.4)
breast cancer 3	0 (0)	0 (0)	0 (0)	**2.8 (0.6)**	5.8 (0.9)	4.2 (0.7)
breast cancer 4	0 (0)	0 (0)	0 (0)	**3.5 (0.7)**	5.3 (1.4)	3.9 (0.7)
ionosphere 1	0.5 (0.2)	0 (0)	2.1 (1.0)	**13.4 (1.8)**	13.7 (1.7)	17.9 (1.9)
ionosphere 2	0 (0)	0 (0)	1.6 (1.0)	**14.8 (1.9)**	18.0 (1.9)	18.0 (2.0)
ionosphere 3	0.6 (0.3)	0 (0)	7.7 (2.0)	**13.1 (1.0)**	14.2 (1.3)	17.1 (3.3)
ionosphere 4	0.7 (0.2)	0 (0)	1.9 (0.2)	**12.0 (1.3)**	14.8 (1.6)	19.1 (1.8)
Pima 1	0.2 (0.1)	0 (0)	0 (0)	23.8 (1.5)	25.4 (1.6)	**23.7 (1.7)**
Pima 2	0 (0)	0.3 (0.2)	0.3 (0.2)	**23.9 (2.2)**	25.3 (0.9)	25.8 (1.5)
Pima 3	0.2 (0.1)	0.1 (0.1)	0.9 (0.4)	**24.0 (1.8)**	25.4 (1.4)	25.6 (1.3)
Pima 4	0.4 (0.2)	0 (0)	0 (0)	**23.7 (1.1)**	24.7 (1.5)	25.4 (1.1)
sonar 1	0 (0)	0 (0)	0 (0)	**24.6 (2.6)**	27.4 (1.9)	27.9 (3.2)
sonar 2	0 (0)	0 (0)	0 (0)	**25.0 (2.4)**	29.3 (1.6)	27.4 (2.4)
sonar 3	0 (0)	0 (0)	0.1 (0.1)	**23.6 (2.4)**	27.3 (4.2)	26.4 (3.3)
sonar 4	0 (0)	0 (0)	0 (0)	26.5 (2.7)	31.7 (2.3)	**24.0 (2.3)**
mean	0.16 (0.07)	0.03 (0.02)	0.91 (0.31)			

Source: own computations.

We also compared ranking based wrapper with *elastic net* to popular stepwise regression which was performed in two directions: forward selection (FS) and backward elimination (BE). The results are summarised in the Table 3. Ranking based wrapper better identified noisy variables and 11 times in 16 datasets led to lower classification error.

Table 3. Mean numbers of noisy variables introduced into the models and classification errors estimated via 10-fold cross-validation (standard errors in brackets)

Dataset	**# noisy variables**			**cv errors**		
	EN	FS	BE	**EN**	FS	BE
breast cancer 1	0 (0)	0.1 (0.1)	3.1 (0.3)	**2.8 (0.8)**	3.2 (0.7)	5.3 (1.1)
breast cancer 2	0 (0)	0 (0)	4.0 (0.5)	**2.6 (0.6)**	4.0 (0.7)	4.9 (0.8)
breast cancer 3	0 (0)	0.1 (0.1)	4.7 (0.5)	**2.8 (0.6)**	4.9 (0.9)	6.2 (1.6)
breast cancer 4	0 (0)	0 (0)	3.9 (0.4)	**3.5 (0.7)**	4.6 (0.8)	5.6 (0.8)
ionosphere 1	0.5 (0.2)	1 (0)	3.4 (0.7)	13.4 (1.8)	13.4 (1.2)	**12.2 (2.0)**
ionosphere 2	0 (0)	0.3 (0.2)	1.8 (1.0)	14.8 (1.9)	14.5 (1.8)	**12.8 (1.3)**
ionosphere 3	0.6 (0.3)	1.5 (0.2)	10.5 (0.4)	13.1 (1.0)	14.8 (2.1)	**12.5 (1.7)**
ionosphere 4	0.7 (0.2)	1.2 (0.1)	9.6 (2.3)	**12.0 (1.3)**	14.3 (1.7)	14.0 (1.7)
Pima 1	0.2 (0.1)	0 (0)	0 (0)	**23.8 (1.5)**	24.2 (1.7)	24.9 (1.2)
Pima 2	0 (0)	0 (0)	0 (0)	23.9 (2.2)	**23.8 (1.4)**	24.4 (1.6)
Pima 3	0.2 (0.1)	0 (0)	0 (0)	**24.0 (1.8)**	24.6 (1.3)	24.2 (0.6)
Pima 4	0.4 (0.2)	0.1 (0.1)	0.1 (0.1)	**23.7 (1.1)**	24.2 (1.8)	24.2 (1.2)
sonar 1	0 (0)	0.1 (0.1)	2.4 (0.3)	**24.6 (2.6)**	28.0 (2.0)	29.4 (3.0)
sonar 2	0 (0)	0 (0)	3.2 (0.4)	**25.0 (2.4)**	30.2 (3.0)	25.5 (2.0)
sonar 3	0 (0)	0.2 (0.1)	4.4 (0.3)	**23.6 (2.4)**	32.1 (4.6)	26.0 (3.3)
sonar 4	0 (0)	0.4 (0.2)	5.0 (0.4)	26.5 (2.7)	**24.9 (3.0)**	25.0 (2.8)
mean	0.16 (0.07)	0.31 (0.08)	3.51 (0.48)			

Source: own computations.

V. SUMMARY

We suggested applying of regularized linear regression for variable ranking in discrimination task. Having conducted the experiments with various numbers and types of noisy variables, we observe that proposed ranking used in wrapper methodology leads to competitive results in comparison to such sophisticated rankings as ones obtained from ensembles. Promising results concern the classification error as well as discarding noisy variables. Additional advantage from proposed ranking is that it reduces the number of variables before the second step of wrapper algorithm is run. Thus, a smaller number of nested models is learned. It can sufficiently reduce the computational cost in the case of high dimension. Moreover, a wrapper based on proposed ranking better discards noisy variables than a popular stepwise regression which is implemented in more statistical software.

REFERENCES

Blum A.L., Langley P. (1997), Selection of relevant features and examples in machine learning, „*Artificial Intelligence*", vol. 97 no. 1-2, p. 245-271.

Breiman L. (2001), Random forests, *"Machine Learning"*, 45, p. 5-32.

Frank A., Asuncion A. (2010), *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science [http://archive.ics.uci.edu/ml].

Freund Y., Schapire R.E. (1996), Experiments with a new boosting algorithm, *Proceedings of the 13th International Conference on Machine Learning*, Morgan Kaufmann, p. 148-156.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji,* PWN, Warszawa.

Gatnar E. (2008), *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji,* PWN, Warszawa.

Guyon I., Gunn S., Nikravesh M., Zadeh L. (2006), *Feature Extraction: Foundations and Applications,* Springer, New York.

Hastie T., Tibshirani R., Friedman J. (2009), *The Elements of Statistical Learning: Data Mining, Inferance, and Prediction,* 2nd edition, Springer, New York.

Hoerl A.E., Kennard R. (1970), Ridge regression: biased estimation for nonorthogonal problems, „*Technometrics*" 12: p. 55-67.

Kubus M. (2011), On model selection in some regularized linear regression methods, *XXX Konferencja Wielowymiarowa Analiza Statystyczna*, Łódź (to appear).

Kubus M. (2013), Liniowy model prawdopodobieństwa z regularyzacją jako metoda doboru zmiennych, *XXI Konferencja Sekcji Klasyfikacji i Analizy Danych PTS*, Lipowy Most (to appear).

Ng A.Y. (1998), On feature selection: learning with exponentially many irrelevant features as training examples, In *Proceedings of the 15th International Conference on Machine Learning*, p. 404-412, San Francisco, CA. Morgan Kaufmann.

Reunanen J. (2006), Search Strategies, In I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), *Feature Extraction: Foundations and Applications,* Springer, New York.

Tibshirani R. (1996), Regression shrinkage and selection via the lasso, *J.Royal. Statist. Soc. B.*, 58: p. 267-288.

Zou H., Hastie T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B*, 67(2): p. 301-320.

Mariusz Kubus

WYBRANE UWAGI NA TEMAT PODEJŚCIA *WRAPPERS* BAZUJĄCEGO NA RANKINGU ZMIENNYCH

Jednym z podejść do problemu selekcji zmiennych w dyskryminacji lub regresji jest wykorzystanie kryterium oceny jakości modeli budowanych na różnych podzbiorach zmiennych (tzw. *wrappers*). Do wyboru podzbiorów zmiennych stosowane są techniki przeszukiwania (heurystyczne lub stochastyczne). Najpopularniejszym przykładem jest regresja krokowa wykorzystująca strategię wspinaczki. Alternatywne podejście polega na uporządkowaniu zmiennych wg wybranego kryterium, a następnie budowaniu modeli zagnieżdżonych i ich ocenie. Zaawansowanymi narzędziami budowy rankingów są agregowane drzewa klasyfikacyjne. W artykule został zaproponowany konkurujący ranking, który prowadzi do nieco mniejszych błędów klasyfikacji. W studium empirycznym rozważane są zmienne nieistotne metryczne oraz binarne. Przedstawiono też porównanie z popularną regresją krokową.