A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 3(302), 2014

Mariusz Kubus[*]

DISCRIMINANT STEPWISE PROCEDURE

Abstract. Stepwise procedure is now probably the most popular tool for automatic feature selection. In most cases it represents model selection approach which evaluates various feature subsets (so called wrapper). In fact it is a heuristic search technique which examines the space of all possible feature subsets. This method is known in the literature under different names and variants. We organize the concepts and terminology, and show several variants of stepwise feature selection from a search strategy point of view. A short review of implementations in R is given.
Keywords: stepwise procedure, feature selection, model selection.

I. INTRODUCTION

In statistics the most popular term for a stepwise procedure is a *stepwise regression*. This is a method for automatic feature selection. It involves the inclusion or removal of a single variable to or from the model, so as to improve its quality. Stepwise regression term first appeared in the article (Efroymson 1960) and is used in the context of linear models. The same feature selection algorithm is also described under the name *sequential selection* (see i.e. Marill and Green 1963; Reunanen 2006). These articles emphasize that the model used in stepwise procedure need not be linear. Another term referring to the variety of stepwise procedures involving the backward elimination of variables is *recursive feature elimination* (RFE). This term is listed very often in combination with SVM (see i.e. Guyon *et al*. 2002), or Random Forest (i.e. Granitto *et al.* 2006). As we will show in Section IV, RFE algorithm is constructed somewhat differently than classical *stepwise regression* (or *sequential selection*) with feature elimination. The significant difference lies in the number of models built, what considerably decrease computational cost.

The goal of this article is to organize the concepts and terminology associated with discriminant stepwise procedure. We will formulate stepwise procedure in general as a solution of combinatorial optimization problem, and show its several variants. Functions of R program which implement stepwise feature selection as well as example of application will also be presented.

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

II. FEATURE SELECTION AS AN OPTIMISATION PROBLEM

The methods of feature selection are currently classified into three groups: filters, wrappers and embedded methods (see i.e. Guyon *et al.* 2006). Filters evaluate the variables in the pre-processing step, so independently of the model. Wrappers use a model for feature subsets evaluation. Model is learned for various feature subsets to choose the best. This approach is also known as model selection. Note that it is about the choice of the model parameters, and not its structure. For example, setting some parameters of the linear model of zero does not change the fact that the model is still linear. Embedded methods have built-in the mechanism of feature selection. This is the case in classification trees or regularized linear regression. The reason for interest in feature selection methods is not only to obtain simple and easy for interpretation model, but also to increase its predictive ability. In predictive modeling, model is learned on the basis of data collected in the past, so-called training set. Suppose we are given a vector of predictors $\boldsymbol{X} = (X_1,...,X_p)$ and response Y, which can be categorical (discrimination) or metrical (regression). The training set consists of multivariate observations with known values of response Y:

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

The task is to learn a model that would accurately classify (or predict) new observations with unknown responses. When there are noisy variables in the data, they can cause the phenomenon of overfitting the model to the data, and thereby reduce its ability of generalization.

Thus, we are dealing with two spaces, the parameter space $\boldsymbol{T}$ and the space of feature subsets $2^{\boldsymbol{X}}$. Assuming some quality criterion of the model Q we are interested in its optimization. Univariate filters reduce the number of variables in the pre-processing stage and estimation of the model is reduced to the estimation of its parameters. In embedded methods learning algorithm simultaneously performs feature selection and estimates the parameters of the model. In wrapper approach the outer loop is applied to the learning algorithm to search the space $2^{\boldsymbol{X}}$. For different subsets of $\boldsymbol{X}$ a model is constructed and evaluation of this model is also the evaluation of a variable subset. Therefore, the feature selection task can be formulated as: find such an element $S \in 2^{\boldsymbol{X}}$ to function $Q : 2^{\boldsymbol{X}} \to R$ has reached the optimum. Since the space $2^{\boldsymbol{X}}$. is finite we are dealing with combinatorial optimization. In a high dimension checking all possible subsets is impractical. Hastie *et al.* (2009) suggest that even for $p > 40$. Therefore, usually only some subsets are checked and the decision which of them are worth

checking is determined by the search strategy. Figuratively speaking search strategy sets out certain paths in the space of all possible subsets. In this perspective, the stepwise procedure is a heuristic search strategy, so a way to solve combinatorial optimization task. Stepwise methods require examining only few subsets of each size in comparison to all possible subsets. They do not guarantee finding optimal solution but they often give useful results in practice. The main drawback of the stepwise procedure is a tendency to reach local optima. It can be overcome by running the algorithm several times from various starting points.

In algorithmic terms the stepwise procedure is also known under the names of greedy search or hill – climbing. In this general sense, we can point a number of tasks in the data analysis, where this procedure is used: learning a classification tree, or generalized additive model, or linear regression model with a use of least angle regression to list some of them. We have posed the feature selection problem for discrimination or regression but stepwise procedure can be applied also in cluster analysis.

III. STEPWISE PROCEDURE AS A SEARCH

The formulation of the search strategy requires defining the space of the search, and determining the five components in the algorithm. This is summarized in Table 1. The basic variants of the stepwise feature selection depend on the operator defining the state change. Usually one feature at a time is added to or removed from the current state. The way in which states are changed is often called the direction of the search. There are three possible variants. Forward selection when only adding of the feature is possible, and backward elimination when only removing of the feature is possible. The third variant is a combination of both, so called bidirectional search. Note that the first two variants construct nested models. In the case of bidirectional search the variable included to the current set can be removed in the following iterations. It is worth pointing that the term stepwise regression is sometimes used in the meaning of bidirectional procedure (see i.e. Efroymson 1960, Aczel 2005). Kittler (1978) or Kudo and Sklansky (2000) consider adding l and removing r features at a time but the idea did not gain greater popularity. This variety is more computationally expensive and more prone to overfitting. Note however, that in modified version discused in Section IV this approach can be competitive (Nagatani *et al.* 2010). Other components of the search strategy provide opportunities to adapt the method to one's own preferences. They act like additional options. The right choice of them may result in constructing a better model.

The starting points are usually empty set (in forward selection) or all features $\boldsymbol{X}$ (in backward elimination). However, this may be any subset of variables. With expert knowledge, we may want to include specified variables to the model. The feature subset can be also obtained in the previous feature selection process. Finally, a subset of the variables can be selected at random. It is practiced in the case when we want to repeat the algorithm several times in order to avoid a local optimum. In the case of bidirectional search, every starting point described above can be applied.

Table 1. Five components of the search algorithm

General approach	Stepwise feature selection
Definition of the search space (the elements are called the states)	The set of all possible subsets of features 2^X
Starting point	Subset of features
The operator of changing the states (the possible steps between the states)	Adding or removing features to/from the current subset. Adding – *forward selection* Removing – *backward elimination* Both – *bidirectional search*
The quality criterion of the states	Evaluation of the feature subset: quality of the model in *wrapper* approach or some statistics in multivariate filters
Stopping criterion	Optional – implemented in high dimensional problems

Source: own research.

The quality criterion of the states depends on the approach to feature selection. In wrapper methodology it is strictly connected with a model. Note that in addition to the decision regarding criterion, one can also choose the model structure. Universal criterion is the accuracy of the model. To avoid the overfitting, accuracy is estimated with a use of resampling methods (cross-validation, bootstrap). Additionally one standard error rule can be applied. It means choosing the model with lesser number of parameters whose error is no more than one standard error above the error of the best model. In linear models more popular are information criteria or statistical tests. The use of them does not require the validation set or performing resampling. Moreover, they can work more effectively what was shown in the context of regularized linear regression (see Kubus 2013a).

A popular stopping criterion is no improvement in quality assessment. A certain minimum value is often determined, by which the assessment should

be improved. It is usually expressed as a percentage. This approach though intuitive and easy to implement has one major drawback, namely reaching of local optima. Other criteria used are the maximum number of iterations or the maximum running time of the algorithm. One can also specify in advance the satisfactory value of evaluation function to stop the algorithm. As shown in Table 1, all of these techniques are to be applied rather in high dimensional problems.

IV. A VARIANT OF STEPWISE PROCEDURE

Classic backward elimination algorithm starts from the set of all variables, and in every step removes the one that causes the greatest improvement of criterion function. In this way, the current set of variables is reduced iteratively until no variable is available. In every iteration the number of models learned is equal to the number of variables in the current set. This can result in high computation cost when firstly p is large and, secondly, when a model itself requires intensive computation, for example, SVM or Random Forests. Guyon *et al.* (2002) proposed *recursive feature elimination* algorithm (RFE) in combination with SVM. This algorithm is similar to stepwise backward elimination, but the key difference is that in each iteration, a model is learned once. This significantly reduces the cost of computation, which is of particular importance in such complex learning algorithms like SVM. The criterion function used in the stepwise procedure is:

$$W^2(\boldsymbol{\alpha}) = \sum \alpha_k \alpha_l y_k y_l K(\boldsymbol{x_k}, \boldsymbol{x_l}), \tag{2}$$

which is inversely proportional to the margin in SVM. When a model is learned, the $\boldsymbol{\alpha}$ coefficients are known, and the criterion is computed p' times by removing from the vectors $\boldsymbol{x_k}, \boldsymbol{x_l}$ one coordinate (p' is the number of variables in the current set). The variable which minimizes the criterion (2) is removed from the current set. Thus, the learned model gives the variables importance but it is recalculated in every iteration. In high dimension more than one variable can be removed in every iteration. Nagatani *et al.* (2010) proposed some modification where more than one variable can be added or removed. We would like to note that the popular linear regression strategy of removing the variables corresponding to the insignificant coefficients and re-building the model, can be formulated as a type of RFE algorithm.

V. R FUNCTIONS AND EXAMPLE

There is no a special package in R for stepwise feature selection. Various functions dedicated to the selected model structures are available in several packages. The short overview is presented in Table 2. Note that RFE algorithm implemented in package `{caret}` can works differently. Default, ranking of features is made only once, what gives ranking based wrapper (see i.e. Kubus 2013b) rather than *recursive feature elimination*. Such an approach does not have much in common with stepwise selection in the classical meaning.

Table 2. Implementations of the stepwise feature selection in R

Functions and packages of R program	Description
`step {stats}` `stepAIC {MASS}`	Stepwise feature selection in three directions (forward, backward and both) with information criterion (for GLMs).
`add1, drop1 {stats}` `addterm, dropterm {MASS}`	Adding or removing a feature controlled by statistical test (for GLMs).
`boot.stepAIC {bootStepAIC}`	A bootstrap procedure under the `stepAIC {MASS}` (for GLMs).
`greedy.wilks {klaR}`	Forward selection with a Wilks lambda criterion (multivariate filter rather than wrapper).
`stepclass {klaR}`	Stepwise feature selection (in three directions). Models can be evaluated using one of the five criteria, i.e. accuracy (estimated via cross – validation). Possible models are: lda, qda, rda, naïve Bayes classifier, kNN, SVM.
`step.plr {stepPlr}`	Stepwise feature selection (forward or both directions) with an information criterion for L2 penalized logistic regression.
`forward.search {FSelector}` `backward.search {FSelector}`	Two directions of greedy search which can be used with any model and evaluation function defined separately.
`hill.climbing.search {FSelector}`	As above but the algorithm starts with a random subset of features.
`rfeIter {caret}` `rfe {caret}`	Recursive feature elimination (after setting the argument `rerank=T` in function `rfeControl`)

Source: own research.

As an example we have used `stepclass` function from package `{klaR}`. Short characteristics of the datasets are given in Table 3. The first two datasets come from UCI Repository of Machine Learning Databases, and the third from R-project documentation for package `{klaR}` (*http://CRAN.R-project.org/package=klaR*).

Table 3. Datasets used in example

Dataset	# observations	# variables	# classes
Ionosphere	351	33	2
Wine	178	13	3
WGBC(*)	157	13	4

*) West German Business Cycles 1955 - 1994

Source: Frank and Asuncion (2010), and R-project documentation for package {klaR}.

Classification error (fraction of incorrectly classified cases) was estimated via 10-fold cross-validation. The method used was kNN with $k = 3$. We run it with and without feature selection. In the first case we applied forward selection. As a model quality criterion we took the accuracy estimated via 10-fold cross-validation. The algorithm was to stop running when the criterion was not improved in some iteration more than 0.1%. The results are presented in Table 4. The use of variable selection in each of the datasets results in a lower classification error. At the same time reducing of the number of variables was also often substantial.

Table 4. The application of stepwise procedure in forward selection variant for k – NN. Classification errors (in %) are estimated via 10-fold cross-validation (standard errors in brackets). The number of selected variables in the cross-validation procedure is given as a range (medians in brackets)

Dataset	No feature selection		Forward selection	
	cv errors	# V	cv errors	# V
Ionosphere	15.08 (2.23)	33	11.39 (1.27)	3-7 (4)
Wine	25.29 (2.75)	13	8.37 (1.89)	2-5 (3)
WGBC	35.83 (3.21)	13	31.87 (2.90)	4-6 (4)

Source: own computations.

VI. FINAL REMARKS

Stepwise feature selection is usually combined with wrapper methodology. The criterion of the feature subset quality is strictly connected with a model, and for this reason wrappers are often preferred over the filters. Stepwise feature selection does not guarantee finding the best solution but it often gives satisfactory results. In high dimension it is widely adopted compromise between

computational complexity and finding the best solutions. This applies especially to forward selection with stopping criterion, which learns the models with a lesser number of variables. In linear models it can be applied even in the case $p > N$. Backward elimination is more computational expensive and a new approach was proposed under the name *recursive feature elimination*. Forward selection is claimed to fail in the case when there are interactions in the data. On the other hand it works quite well in the presence of redundant variables. Backward elimination behaves inversely.

REFERENCES

Aczel A.D. (2005), *Statystyka w zarządzaniu,* PWN.

Efroymson M. A. (l960), Multiple regression analysis. *Mathematical Methods for Digital Computers,* Ralston **A.** and Wilf, H. S., (eds.), Wiley, New York.

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/).

Granitto P.M., Furlanello C., Biasioli F., Gasperi F. (2006), Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products, *Chemometrics and Intelligent Laboratory Systems,* Vol. 83, 2, p. 83–90.

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

Guyon I., Weston J., Barnhill S., Vapnik V. (2002), Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, 46: 389–422

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

Kittler J. (1978), Feature set search algorithms, In C. H. Chen, editor, *Pattern Recognition and Signal Processing*, p. 41–60, Sijthoff & Noordhoff.

Kubus M. (2013a), On model selection in some regularized linear regression methods, In: Domański Cz., Kupis-Fijałkowska A. (Eds.) *Multivariate Statistical Analysis – Theory and Practice*, Acta Universitatis Lodziensis, Folia Oeconomica 285, p. 115–223.

Kubus M. (2013b), Some remarks on feature ranking based wrappers, In: Domański Cz. (Ed.) *Methods and Applications of Multivariate Statistical Analysis*, Acta Universitatis Lodziensis, Folia Oeconomica 286, p. 147–154.

Kudo M., Sklansky J. (2000), Comparison of algorithms that select features for pattern classifiers, *Pattern Recognition*, 33(1):25–41.

Marill T., Green D. M. (1963), On the effectiveness of receptors in recognition systems, *IEEE Transactions on Information Theory*, 9(1): 11–17.

Nagatani T., Ozawa S., Abe S. (2010), Fast Variable Selection by Block Addition and Block Deletion, *Journal of Intelligent Learning Systems and Applications*, Vol. 2 No. 4, p. 200-211. doi: 10.4236/jilsa.2010.24023.

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

Mariusz Kubus

DYSKRYMINACYJNA PROCEDURA KROKOWA

Procedura krokowa jest obecnie chyba najpopularniejszym narzędziem automatycznej selekcji zmiennych. Najczęściej prezentuje podejście polegające na ocenie podzbiorów zmiennych za pomocą kryterium jakości modelu (*wrapper*). W istocie jest techniką heurystycznego przeszukiwania przestrzeni wszystkich podzbiorów oryginalnego zestawu zmiennych. Metoda ta znana jest z literatury pod różnymi nazwami i w różnych wersjach. W artykule przedstawiono ogólny schemat krokowej selekcji zmiennych oraz wskazano różne jej warianty. Uporządkowano terminologię oraz podano krótki przegląd funkcji programu R implementujących krokową selekcję zmiennych.