МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2014
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2014
Proceedings of the Forty Third Spring Conference
of the Union of Bulgarian Mathematicians
Borovetz, April 2–6, 2014

APPLICATION OF GENERALIZED PATHSEEKER REGULARIZED REGRESSION*

Snezhana Gocheva-Ilieva

This paper presents an application of one of the latest regression methods – generalized PathSeeker (GPS), stemming from the new generation of predictive techniques which use the data mining approach. The method is designed for statistical analysis of data, when classical parametric modeling is not applicable or does not provide sufficiently good results. Experimental data from the field of laser technology have been processed using GPS and associated data mining techniques such as TreeNet, ISLE and RuleRunner. The influence of 10 operating laser parameters on the output power of copper bromide vapor lasers has been modeled. The obtained best linear regression models have high coefficients of determination $R^2 = 98\%$ for the learn sample and 97% for the test sample. Advantages and disadvantages of the GPS method have been discussed.

Introduction. Regression is among the most preferred and the most used methods for fitting to data. It is widely used in natural sciences, engineering, environmental, behavioral and social sciences, and other areas to describe possible relationships between variables and build appropriate models. The models can then be applied for prediction or forecasting future observations or for quantifying the strength of the relationship between investigated variables.

The foundations of the classical regression method originate in the writings of Legendre (1805), Gauss (1809) and Fisher (1920). However, new challenges arise with the appearance of huge datasets with high dimensions, multicollinearity, nonlinear interactions between variables, and the need for higher practical accuracy and predictive ability of the constructed models.

Consider a data set of n observations $\{y, \mathbf{X}\} = \{y_i, \mathbf{X}_i\}_{i=1}^n = \{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$, where y is the dependent variable and $\mathbf{X}$ is the matrix of p vectors of independent variables $X_1, X_2, \ldots, X_p$. The classical multiple linear regression (MLR) model for fitting to these data is

$$\hat{y} = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_p X_p, \quad \varepsilon_i = y_i - \hat{y}_i, \qquad i = 1, \ldots, n \tag{1}$$

where $\mathbf{a} = (a_0, a_1, \ldots, a_p)$ is the vector of the regression coefficients (or parameters) and ε_i are the residuals (or error terms) of the model. The model (1) assumes that the dependence between y and $X_1, X_2, \ldots, X_p$ is linear and the residuals form an unobserved

*2010 Mathematics Subject Classification: 62J07, 62P30, 62P35.
Key words: regularized regression, ridge regression, LASSO, TreeNet, copper bromide laser.
This paper is supported by the project NI13–FMI-002, NPD, Paisii Hilendarski University of Plovdiv.

random variable. The dependent variable y is also called response and the independent variables are called predictors or regressors. The problem is to determine the estimates $\hat{\mathbf{a}}$ of the regression coefficients which minimize a given loss function.

Usually the loss function is taken as a mean squared error (MSE)

$$L(y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n. \tag{2}$$

In this case a unique solution exits, known as ordinary least squared (OLS) regression model.

In practice, the application of classical regression exhibits many problems. The main of these are: the fulfilling of the theoretical assumptions of the method, including the requirement of the normal distribution of the response, determination of the substantial predictors to include in the model, lack of external knowledge (because all data are used for the model construction), appearance of unstable solutions in the case of multicollinearity and more.

To overcome these problems many other types of regression methods and techniques, parametric and nonparametric ones have been developed. An overview of the current state and the capabilities of regression and predictive statistical techniques, including of data mining methods, can be found, for example, in [1–3].

Alternative approaches to constructing useful regression models based on linear combinations of the form (1) include the well-known ridge regression, developed in [4], and the least absolute shrinkage and selection operator (Lasso) regression [5], as well as their generalizations as elastic net family [6], and other methods. These methods are based on the use of the penalized coefficient estimation. This usually introduces a bias into the estimation, but leads to reduced variability of the estimate.

One of the latest regression methods is the generalized PathSeeker (GPS) regularized regression, entirely oriented towards algorithmization and extensive computer calculations. GPS is a new generation statistical method, developed by Jerome Friedman in 2008 [7] and implemented in the spring of 2013 as part of the software package Salford Predictive Modeler [8, 9]. Despite its high performance, as a linear regression technique, the GPS method is sensitive to the distribution of data. To improve its applicability, GPS is used in combination with additional techniques such as TreeNet, known also as ensemble of boosted trees, Importance Sampled Learning Ensembles (ISLE) and RuleLearner [10–12].

In this paper the GPS is applied to construct and analyze an empirical model in the area of laser physics. Experimental data have been studied for a family of copper bromide (CuBr) lasers, developed in the Georgi Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences [13]. Based on these data, various statistical models were built in [14] by using multiple linear regression, principal component regression, nonlinear regression and multivariate adaptive regression splines (MARS). In the recent papers [15, 16] high quality models have been built and examined using the MARS and CART (Classification and Regression Trees) methods [17, 18].

In this study we first construct a GPS model by using the initial variables similarly to classical regression. The second type of model is constructed on the basis of the ensemble of the trees, generated by TreeNet. For the compression of the trees and post-processing additional techniques such as ISLE and RuleLearner have been applied. A

comparison between the best GPS models obtained and the classical stepwise multiple linear regression model is presented.

The GPS models are built by means of the Salford Predictive Modeler software [8] and the stepwise regression model is obtained by using the SPSS statistical package [19].

1. Brief introduction of the regularized regression methods and GPS regression. The standard regularized modeling aims to find a regression equation (1) for fitting to data by solving the following optimization problem for the regression coefficients

$$\hat{\mathbf{a}}(\lambda) = \underset{\mathbf{a}}{\operatorname{argmin}} \left[\widehat{R}(\mathbf{a}) + \lambda P(\mathbf{a})\right] \tag{3}$$

where $R(\mathbf{a})$ is the empirical loss function, selected among different error criteria (e.g. MSE in (2)), $P(\mathbf{a})$ is a penalty function and $\lambda > 0$ is the regularization parameter. In ridge regression [4] the penalty function is selected as $P(\mathbf{a}) = \sum_{j=0}^{p} a_j^2$. The other well-known method is the Lasso method, which uses $P(\mathbf{a}) = \sum_{j=0}^{p} |a_j|$. In the extended power family [7] the penalty function is generalized to the form $P(\mathbf{a}) = P_\gamma(\mathbf{a}) = \sum_{j=0}^{p} |a_j|^\gamma$, $0 \le \gamma \le 2$. For $\gamma = 1$ one obtains the above mentioned Lasso regression, which allows to introduce variables sparingly and generate reasonably sparse solutions. For $\gamma = 2$ it is Ridge regression, which allows the sparing introduction of variables and generates reasonably sparse solutions. The Ridge regression contributes to estimating stabilization in the presence of extreme multicollinearity. The difference between Lasso and ridge regression is that in ridge regression, as the penalty is increased, all parameters are reduced while still remaining non-zero, while in Lasso, increasing the penalty will cause more and more of the parameters to be driven to zero.

Further extension in terms of penalty function is represented by the elastic net family as the combination [6]:

$$P_\beta(\mathbf{a}) = \sum_{j=1}^{p} (\beta - 1)\, a_j^2/2 + (2 - \beta)\, |a_j|, \quad 1 \le \beta \le 2, \tag{4}$$

where β is the coefficient of elasticity, which is further extended with a suitable formula for $0 \le \beta < 1$ [7].

Other similar methods have been proposed for different penalties in (3), along with minimization methods. Algorithmically, the model coefficients are found by seeking the minimum (3), for example by using the gradient descend method (see [7]).

With the GPS method, the abovementioned methods are generalized and the problem of calculation complexity has been resolved through sequential path search directly in the parameter space under a given penalty $P(\mathbf{a})$ without solving the optimization problem at each step. Especially, for all $\mathbf{a}$ the penalty function has to satisfy the condition

$$\left\{ \frac{\partial P(\mathbf{a})}{\partial\, |a_j|} > 0 \right\}_1^p. \tag{5}$$

The class (5) includes as a special case the families of power and elastic net penalties, as well as all known penalties of the same type. The meta-logic of the GPS algorithm and illustrative examples with analysis are given in [7].

The GPS regularized regression is designed to handle continuous or binary data and builds high-quality linear models in the usual form (1) producing a number of paths (classes) of regression. GPS is implemented in Salford Predictive Modeler software as a very fast forward stepping algorithm with a specialized variable selection procedure

[8]. There is generated a collection of models by constructing a path based on selected predictors $\mathbf{X}$ as a sequence of iterations (steps) in the space of coefficients. This includes zero coefficient model, sequence of 1-variable models, sequence of 2-variable models, etc. At every step a new variable selected to fulfill a complex of criteria is added, or the coefficient of some model variable is adjusted. The quality of models is achieved using a number of commonly used goodness-of-fit measures for learn and test samples as the coefficient of determination R^2, MSE, etc.

Despite its advantages, the GPS has some limitations [7]. As a linear regression technique it is sensitive to data distribution. The method does not provide automatic discovery of nonlinearities, interactions between predictors, or a missing values handling feature. In practice, the use of GPS as a data mining engine is highly efficient in combination with model simplification techniques mentioned above. The basic one among them is TreeNet stochastic gradient boosting, which is applied for preprocessing the data in order to obtain more adequate predictors, based on the generated trees. This technique is designed to handle both regression and classification problems. The original model, produced by TreeNet is an ensemble of hundreds or even thousands of small trees $T_i(\mathbf{X})$ (normally with 2 to 6 terminal nodes) in the form

$$\tilde{y} = y^0 + b_1 T_1(\mathbf{X}) + b_2 T_2(\mathbf{X}) + \cdots + b_M T_M(\mathbf{X}). \tag{6}$$

where M is the number of trees. Usually, for the least square loss criterion (2), $y^0 = \bar{y}$ (mean value of y), which gives an initial model. Then using this response all the residuals $g_i^0 = y_i - y_i^0$, $(i = 1, 2, \ldots, n)$ are computed. At any current stage of construction of the next tree, respectively the next model, a random sample is drawn from the learn data and the tree built by using current residuals from the response as dependent variables. The original TreeNet model combines all trees with equal coefficients. A more detailed description of the algorithm is given in [10]. In essence every tree can be considered as a new variable transformation represented by a continuous variable as a function of inputs. Furthermore, every node (internal or terminal) can also be represented as a dummy variable. These variables are then used by GPS as predictor variables.

In fact, many of the obtained trees are usually equal or have very similar structure. The compression of the TreeNet model can be performed using the ISLE algorithm by removing redundant trees. The coefficients of the models (6) are then adjusted by the GPS method. The RuleLearner algorithm has to be applied as a post-processing technique which selects the most influential subset of nodes, thus further reducing model complexity. Different combinations of these methods can also be carried out to obtain the best model for fitting to data.

2. Description of experimental data for CuBr laser. We will model the data of CuBr laser. The development of this type of lasers continues to be topical [13]. This is due to the fact that in the visible range (wavelengths 510.6 nm and 578.2 nm), these lasers operate at highest output power and have unique properties. One of the important technological objectives is the development of new laser devices of this type with enhanced output power. In particular, statistical modeling supports the investigation of the influence of the main laser operating parameters (laser tube geometry, input power, neutral gas pressure, etc.) on output laser power and allows predictions to be made.

We consider data of 10 input basic variables which determine the CuBr laser operation: D (mm) – inner diameter of the laser tube, DR (mm) – inner diameter of the rings, L

(cm) – electrode separation (length of the active area), PIN (kW) – input electrical power, PRF (KHz) – pulse repetition frequency, PNE (Torr) – neon gas pressure, $PH2$ (Torr) – hydrogen gas pressure, PL (kW/cm) – specific electrical power per unit length, C (nF) – equivalent capacity of the condensation battery, TR (°C) – temperature of CuBr reservoirs. The response variable is $Pout$ (W) – output laser power. The data are of historical type. The sample size is $n = 387$. Here we have to mention the complexity, long duration and high cost of each conducted experiment. A detailed description of the laser device, the data and their sources are given in [13, 14].

It can be mentioned that although the general population is assumed to be normally distributed, the available data, and especially the response variable $Pout$, is not normally distributed. Also high multicollinearity exists between some of the independent variables [14]. This way, the statistical analysis of the considered data set requires application of appropriate statistical techniques to direct the experiments aimed at further development of the copper bromide laser devices.

3. Construction of GPS model using the 10 initial predictors. We first construct the model by applying the GPS method with the initial 10 independent variables and the dependent variable $Pout$. When desiring to build a model, we need to set up the basic control parameters of the selected analysis method. The model setup procedure includes the selection of elasticities (given in the brackets): Compact (0.0), Lasso (1.0), Ridged Lasso (1.1) and Ridge regression (2.0), where the Compact method with elasticity 0.0 is similar to the stepwise multiple regression. There can be selected the volumes of the learn and test samples, penalty constrains, different regression performance evaluation criteria, path controls as a maximum number of the steps and a maximum number of the points in the path and others. As a rule, all variables are standardized before processing to equalize their influence in the model. Also, if necessary, some procedures for handling outliers and missing values could be adjusted. The confidence bounds for the best model are usually established at level 0.05 or at another given level.

For our model we choose the analysis method GPS with all types of penalties, MSE and the coefficient of determination R^2 as main regression performance criteria, and a significance level of 0.05. The learn sample is chosen as 80% or 319 randomly selected cases and the test sample is the remaining 20% or 68 of the whole investigated sample. Other controls are stated by default.

The obtained best GPS model is Lasso (1.0) or GPS with elasticity $\beta = 1$. The coefficients of the best model are given in the second column of Table 1.

It can be observed that only 4 coefficients are substantially non-zero, the remaining 6 coefficients are small and are taken to be zero in the optimal model. In the 3$^{\text{rd}}$ column are given the values of the variable importance of the four non-zero coefficients in the model. The input electric power PIN has the biggest influence, followed by C, $PH2$ and PL. This result is in very good agreement with other type of statistical models, obtained for instance in [15, 16]. In the last two columns there are given the coefficients of the classical stepwise multiple linear regression and their significance. Excluding the insignificant constant and excluded variable $PH2$, the values of other available regression coefficients are similar to these in the optimal GPS model. We have to give here the coefficient of determination of the classical stepwise MLR model, which is $R^2 = 0.931$ (see also Fig. 1a)).

The main selected criteria for model performance are the coefficient of determination

Table 1. The coefficients and some characteristics for models with 10 initial variables

Variable	Best GPS model: Lasso(1.0)		Classical stepwise regression	
	Coefficients	Variable importance[a]	Coefficients	Significance
Constant	−25.41652	–	−14.160	0.104
PIN	10.98233	100	11.044	0.000
L	0.21793	–	0.219	0.000
DR	0.33317	–	0.460	0.000
$PH2$	5.43409	46.7	–	–
C	−6.14385	53.3	−6.258	0.000
PRF	−0.03769	–	–	–
TR	−0.03265	–	−0.049	0.001
D	0.13923	–	–	–
PL	1.63285	13.3	1.580	0.003
PNE	−0.01892	–	–	

[a]The values of the variable importance is relative to the biggest importance, equal to 100.

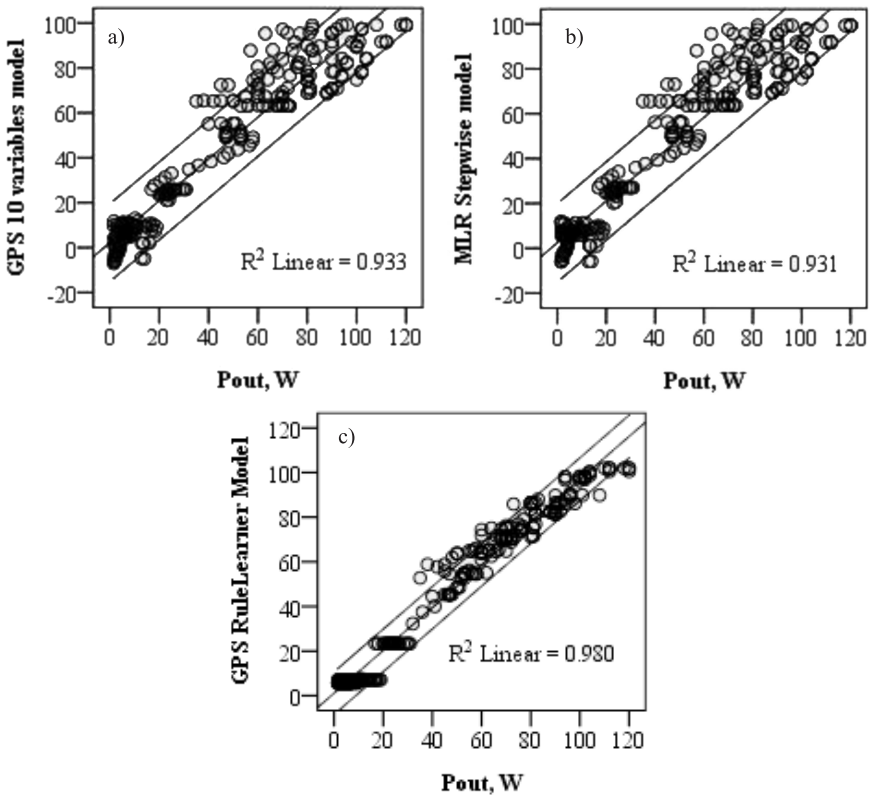

Fig. 1. Comparison of the experimental data of the output laser power Pout versus the predicted values from the following models: a) stepwise multiple linear regression; b) the best GPS model, based on the 10 initial independent variables; c) the best GPS/Rule Learner – ISLE model

R^2 and MSE, for the learn and test samples. Other standard measures of goodness-of-fit such as RMSE (root MSE), MAD (mean absolute deviation), AIC (Akaike Information Criterion) and others are also calculated and can be used for model evaluation. The basic performance characteristics of the obtained GPS models in this study are given in Table 2. The plots of the model predicted values against the experimental values of the output laser power *Pout* are given in Fig. 1a) and b).

4. Construction of GPS model using the variables from the TreeNet tree ensemble. As was mentioned above, the GPS regression has some limitations that could be overcome by applying it jointly with the TreeNet and other data mining techniques. The idea is to preprocess the initial data set by the TreeNet method, which produces regression models in the form of a large amount of small decision trees (6) [10]. Each tree typically contains about six terminal nodes. On the basis of the trees new predictor variables are determined. This way a new data set can be obtained in order to run a regularized regression, namely the GPS, and to improve the model. As was mentioned above, the ISLE and Rule Learner techniques are designed to optimize the TreeNet original solution by compressing the tree ensemble and post-processing the rules.

Table 2. Basic statistics of model performance for the obtained optimal GPS models

Criterion	Best GPS 10-variables model: Lasso(1.0)		Best GPS/RuleLearner model: ISLE – Lasso(1.0)	
	Learn sample	Test sample	Learn sample	Test sample
MSE	80.471	94.211	21.805	43.488
RMSE	8.971	9.706	4.670	6.595
MAD	5.978	6.974	3.247	4.620
MRAD	0.380	0.415	0.272	0.260
R^2	0.933	0.930	0.982	0.968
AIC	1421.742	331.098	1019.209	292.530

In our analysis we used all possible combinations of all these techniques and carried out the GPS regularized regression by choosing the upper limit of 200 trees. All other control parameters were fixed to the same settings as in the previous analysis in Section 3. For the original TreeNet model, the MSE as a function of the number of trees is shown in Fig. 2.

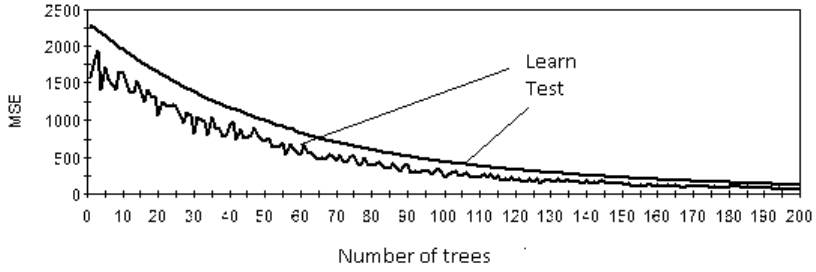

Fig. 2. Dependence of the MSE on the number of trees in the TreeNet model

One can observe the fast improvement in terms of the mean squared error when the

number of trees increases.

The model performance for different methods is illustrated in Fig. 3. Detailed information is given in Table 3. Here the ISLE model with GPS-Lasso(1.0) is the optimal one, with only 17 trees included in the model, respectively 18 non-zero coefficients, taking into account the constant term. The basic statistics of model performance are presented in the last 2 columns of Table 2. The comparison of the predicted values with the experimental values of *Pout* are plotted in Fig. 1c).

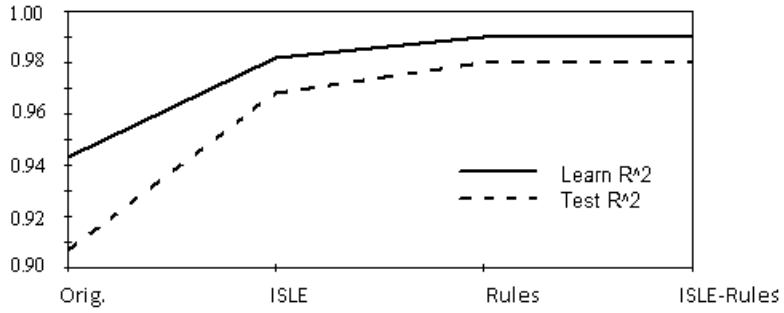

Fig. 3. Model performance of the RL techniques. The noted original model corresponds to the pure TreeNet model

Table 3. Basic statistics of model performance for the different GPS models

Model	Learn R^2	Learn N Coef.	% Compression	Test R^2	Test N Coef.	% Compression	Elasticity
Original (TreeNet)	0.94263	195	0.0%	0.90712	200	0.0%	
ISLE	0.98172	17	91.3%	0.96762	17	91.0%	Lasso (1.0)
RuleLearner	0.98994	81	95.8%	0.98014	81	95.5%	Lasso (1.0)
ISLE_RuleLearner	0.98993	83	95.7%	0.98013	83	95.5%	Lasso (1.0)

5. Discussion with conclusion. From Table 1 and Fig. 1a) and b) one can see that the best GPS model based only on the initial 10 variables and the response has almost the same performance as the classical stepwise MLR model. As shown in Table 2, the obtained best GPS models demonstrate very good statistical indices, including relatively high values of the coefficient of determination R^2 of around and over 93%. The second best GPS/RuleLearner model was found to be an ISLE model with R^2 equal to 98% for the learn sample and 97% for the test sample. This performance is in good agreement with 95% accuracy of the physical measures. In conclusion, all derived models can be used to describe and predict the experiment.

At the same time, from a practical point of view, the predictions of the models in the region of high laser output power as seen in Fig. 1, are not satisfactory and needs specific model selection strategy. This can be based on using special holdout samples when checking the model performance criteria. Also appropriate preliminary transformations of data or other approach may improve the accuracy of the models.

Finally, we can note that the constructed models are comparable to the other existing models built on the same data sample using CART and MARS methods [15, 16]. It can be concluded that the obtained models can be further analyzed and used in order to improve the output laser characteristics, in our case the very important one – the output laser power.

REFERENCES

[1] NIST/SEMATECH e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook/

[2] R. Nisbet, J. Elder, G. Miner. Handbook of statistical analysis and data mining applications. Burlington: Elsevier Academic Press, 2009.

[3] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer, 2001.

[4] A. E. Hoerl, R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **42**, *1* (1970), 80–86.

[5] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statistical Soc.*, Ser. B, **58**, *1* (1996), 267–288.

[6] H. Zou, T. Hastie. Regularization and variable selection via the elastic net. *J. Royal Statist. Soc.*, Ser. B, **67** (2005), 301–320.

[7] J. H. Friedman. Fast sparse regression and classification. Stanford University Technical Report, 2008; published also in: *International Journal of Forecasting*, **28**, *3* (2012), 722–738.

[8] http://salford-systems.com

[9] Salford Predictive Modeler software – Users' Guide. Introducing Generalized PathSeeker. San Diego, Salford Systems, 2013.

[10] J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. 1999 Reitz Lecture. *Annals of Statistics*, **29**, *5* (2001), 1189–1232.

[11] J. H. Friedman, B. E. Popescu. Importance sampled learning ensembles. Stanford University, Department of Statistics. Technical Report, 2003.
http://www-stat.stanford.edu/~jhf/ftp/isle.pdf

[12] J. H. Friedman, B. E. Popescu. Predictive Learning Via Rule Ensembles, 2003.
http://www-stat.stanford.edu/~jhf/ftp/RuleFit.pdf

[13] N. V. Sabotinov. Metal vapor lasers. In: Gas Lasers (Eds M. Endo, R. F. Walter). Boca Raton, CRCPress, 2006, 449–494.

[14] S. G. Gocheva-Ilieva, I. P. Iliev. Statistical Models of Characteristics of Metal Vapor Lasers. New York, Nova Science Publishers Inc., 2011.

[15] I. P. Iliev, D. S. Voynikova, S. G. Gocheva-Ilieva. Simulation of the output power of copper bromide lasers by the MARS method. *Quantum Electronics*, **42**, *4* (2012), 298–303.

[16] I. P. Iliev, D. S. Voynikova, S. G. Gocheva-Ilieva. Application of the classification and regression trees for modeling the laser output power of a copper bromide vapor laser. *Mathematical Problems in Engineering*, **2013**, (2013), Article ID 654845, 1–10.

[17] J. H. Friedman. Multivariate adaptive regression splines (with discussion), *The Annals of Statistics*, **19**, *1* (1991), 1–141.

[18] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. Classification and Regression Trees. Belmont, Wadsworth Advanced Books and Software, 1984.

[19] SPSS IBM Statistics, 2013. http://www-01.ibm.com/software/analytics/spss/

Snezhana Gocheva-Ilieva
Faculty of Mathematics, Informatics and Information Technology
Paisii Hilendarski University of Plovdiv
24, Tsar Assen Str.
4000 Plovdiv, Bulgaria
e-mail: snow@uni-plovdiv.bg, snegocheva@gmail.com

ЕДНО ПРИЛОЖЕНИЕ НА ОБОБЩЕНАТА PATHSEEKER РЕГУЛЯРИЗИРАЩА РЕГРЕСИЯ

Снежана Гочева-Илиева

Тази работа представя едно приложение на един от най-новите регресионни методи Generalized PathSeeker (GPS), отнасящ се към новата генерация предсказващи статистически техники, основаващи се на подхода за интелигентна обработка на данни. Методът е разработен за статистически анализ на данни, когато класическото параметрично моделиране не е приложимо или не довежда до достатъчно добри резултати. С помощта на GPS и асоциираните с него техники като TreeNet и RuleLearner са обработени експериментални данни от областта на лазерните технологии. Построени са модели за моделиране на влиянието на 10 лазерни работни параметри върху изходната мощност на лазери с пари на меден бромид. Получените най-добри линейни регресионни модели имат много добро качество, с коефициенти на детерминация $R^2 = 98\%$ за обучителната и 97% за тестовата извадка. Дискутирани са преимуществата и недостатъците на метода.