CORRELATION MAXIMIZATION IN REGRESSION MODELS BASED ON CONVEX COMBINATIONS

Oleg Senko, Alexander Dokukin

Abstract: A new regression method based on convex correcting procedures over sets of predictors is developed. In contrast to previously developed approach based on minimization of generalized error, the proposed one utilies correcting procedures of maximal correlation with the target value. In the proposed approach a concept of a set of predictors irreducible against target functional is used where irreducibility is understood as lack of combinations of at least the same value of the functional after removing any of its predictors. Sets of combinations simultaniously irreducilbe and unexpandable are used during the construction of a prognostic rule. Results of some computational experiments described in the present article show an efficiency comparison between the two approaches.

Keywords: forecasting, bias-variance decomposition, convex combinations, variables selection.

ACM Classification Keywords: G.3 Probability and Statistics - Correlation and regression analysis, Statistical computing.

Introduction

Several statistical methods were developed last years that allow improving significantly prognostic ability of regression modeling in tasks of high dimension. Efficiency of these methods is associated with effective selecting of prognostic variables. Least angle regression or Lasso [Efron et al., 2004], [Tibshirani, 1996] methods may be mentioned thereupon. However a problem of low generalization ability of empirical models in high-dimensional tasks cannot be considered completely solved. Development of new alternative approaches may be useful for estimating of forecasting ability upper boundaries or for evaluating of selected variables optimal number. An approach in which optimal forecasting models are built by ensembles of preliminary trained predictors is discussed in this paper. It is supposed that initial predictors are simple. For example they may be one-variate or two-variate regression models. Suppose that we have set of L predictors $z_1,\dots,z_L$ that forecast some variable Y. Let $c=(c_1,\dots,c_L)$ be a vector of nonnegative coefficients satisfying condition $\sum_{i=1}^{L} c_i = 1$. Convex correcting procedure (CCP) calculates forecasted value as a weighted sum of prognoses that are calculated by single predictors:

$$Z_{ccp}(c)=\sum_{i=1}^{L} c_i z_i \ .$$

Convex combinations are widely used in pattern recognition. The bagging and boosting techniques [Breiman, 1999], [Kuncheva, 2004] may be mentioned as an example, as well as methods based on collective solutions by sets of regularities [Zhuravlev et al., 2008], [Zhuravlev et al., 2006], [Kuznetsov et al., 1996]. Convex correction is used in regression tasks also. Thus, neural networks ensembles are discussed in [Brown et al., 2005] that are based on optimal balance between individual forecasting ability of predictors and divergence between them. Efficiency of convex combinations of repressors’ pairs was shown in [Senko, 2004]. Earlier it was shown that error of predictors’ convex combination in any case is not greater than the same convex combination of single predictors’ generalized errors [Krogh et al., 1995]. In previous works [Senko, 2009], [Senko et al., 2010] a method for CCP optimization has been studied that is based on minimization of general error estimates. Experiments with simulated data demonstrated that CCP error optimization also implements effective selection of informative prognostic variables. It is easy showing that the decrease of CCP variance comparing to the same combination of single predictors’ variances is also a quality of convex combinations. Such a decrease deteriorates the CCP’s prognostic ability. So, CCP predictions must be additionally adjusted, that may be done with the help of simple linear uni-variate regression. But forecasting ability of a linear regression model depends monotonically on correlation coefficients between Z_{ccp} and Y . In this paper we develop a new technique for constructing the CCP of maximal correlation with Y . This technique is based on the same concept of irreducible ensembles searching that was used in [Senko et al., 2010].

It is supposed further that predictors from initial set are additionally transformed with the help of optimal uni-dimensional regression models to achieve best forecasting ability. Such predictors will be further called reduced. In other words predictor z will be called reduced if for all α, β the inequality

$$E_{\Omega}\left(Y-\alpha z-\beta\right)^{2} \leq E_{\Omega}\left(Y-z\right)^{2}$$

is correct. Here $E_{\Omega}\left(X\right)$ is mathematical mean of X by space of admissible objects with defined σ-algebra and probability measure. It will be further denoted as $\hat{X}$. Variance of X will be denoted as $V(X)$. It is known that following equalities are true for a reduced predictor z :

$$\operatorname{cov}\left(Y,z\right)=E_{\Omega}\left[\left(Y-\hat{Y}\right)\left(z-\hat{z}\right)\right]=E_{\Omega}\left(z-\hat{z}\right)^{2}.$$

The use of the described conditions allows effectively searching ensembles with maximal prognostic ability, but the approach has its drawbacks. First of all, there are many ensembles with the prognostic ability close to the optimal one and it would be rational using them all. Secondly, CCP always decrease prognoses’ variation and uni-dimensional correcting transformation becomes inevitable. Of all predictors the maximal quality is provided by the one most correlated with Y.

Irreducible ensembles relatively correlation coefficients

Standard Pearson correlation coefficient is defined as the ratio:

$$K\left(Y,Z_{ccp}\right)=\frac{\operatorname{cov}\left(Y,Z_{ccp}\right)}{\sqrt{V\left(Y\right)V\left(Z_{ccp}\right)}}.$$

On the other hand $\operatorname{cov}\left(Y,Z_{ccp}\right)=\sum_{i=1}^{L}c_i\operatorname{cov}\left(Y,z_i\right)$. But z_i is a reduced predictor. So, $\operatorname{cov}\left(Y,z_i\right)=V\left(z_i\right)$, $i=1,\dots,L$ and therefore

$$K\left[Y,Z_{ccp}\left(c\right)\right]=\frac{\sum_{i=1}^{L}c_iV\left(z_i\right)}{\sqrt{V\left(Y\right)}\sqrt{\sum_{i=1}^{L}c_iV\left(z_i\right)-\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}c_ic_j\rho_{ij}}}.$$

Further discussions are based on irreducible ensemble concept. A set of predictors $\tilde{z}$ is called irreducible ensemble if removing of at least one predictor from it does not allow constructing CCP with the same prognostic ability as of $\tilde{z}$. The following is a strict definition of ensemble’s irreducibility.

Definition 1. Sets $\overline{D_L}$, D_L from $\mathbb{R}^L$ are defined as

$$\overline{D_L}=\left\{c\left|\sum_{i=1}^{L}c_i=1;c_i\geq 0,i=1,...,L\right.\right\},$$

$$D_L=\left\{c\left|\sum_{i=1}^{L}c_i=1;c_i>0,i=1,...,L\right.\right\}.$$

Definition 2. Set of predictors $z_1,...,z_L$ is called irreducible ensemble relative to some functional $F\left(c\right)$, that characterize forecasting ability, if there is such vector $c^*\in D_L$, that $\forall c'\in\overline{D_L}\setminus D_L$, $F\left(c^*\right)>F\left(c'\right)$.

A set of points from $\mathbb{R}^L$ simultaneously satisfying constraints: $\sum_{i=1}^{L}c_i=1$ and $\sum_{i=1}^{L}c_iV\left(z_i\right)=\theta$ will be further referred to as $W\left(\theta\right)$.

Theorem 1. A necessary condition of irreducibility of predictors set $z_1,...,z_L$ relative to $K\left(Y,Z_{ccp}\right)$ is existence of such real θ that quadratic functional

$$P_f\left(c\right)=\sum_{i=1}^{L}\sum_{j=1}^{L}c_ic_j\rho_{ij}^{v}.$$

achieves strict maximum at $W\left(\theta\right)$ in $c_1^*,...,c_L^*$ that satisfies conditions $c_i^*>0$, $i=1,...,L$.

The maximum necessary condition is existing of positive $\theta>0$, such that the following equation holds

$$\sum_{i=1}^{L}\sum_{j=1}^{L}c_ic_j\rho\left(z_i,z_j\right)\rightarrow\max \tag{1}$$

with the next contingencies:

$$\sum_{i=1}^{L}c_iV\left(z_i\right)=\theta,$$

$$\sum_{i=1}^{L}c_i=1,$$

$$c_i\geq 0,\ i=1,...,L. \tag{2}$$

Lets write down a Lagrange functional for the task (1)

$$L=\sum_{i=1}^{L}\sum_{j=1}^{L}c_ic_j\rho\left(z_i,z_j\right)+\lambda\left(\sum_{i=1}^{L}c_iV\left(z_i\right)-\theta\right)+\mu\left(\sum_{i=1}^{L}c_i-1\right),$$

and equal its partial derivatives to zero

$$\frac{\partial L}{\partial c_k}=2\sum_{i=1}^{L}c_i\rho\left(z_i,z_k\right)+\lambda V\left(z_k\right)+\mu=0,$$

$$\frac{\partial L}{\partial \lambda}=\sum_{i=1}^{L}c_iV\left(z_i\right)-\theta=0,$$

$$\frac{\partial L}{\partial \mu}=\sum_{i=1}^{L}c_i-1=0.$$

Moving to a vectorial form we get

$$2DC+\lambda E+\mu I=O,$$

$$E^TC=\theta,$$

$$I^TC=1.$$

where $D=\left\|\rho\left(z_i,z_j\right)\right\|_{n\times n}$, $E=\left\|E\left(z_i^2\right)\right\|_{1\times n}=\left\|V\left(z_i\right)\right\|_{1\times n}$, $C=\left\|c_i\right\|_{1\times n}$, $I=\left\|1\right\|_{1\times n}$, $O=\left\|0\right\|_{1\times n}$.

Lets denote $\alpha=E^TD^{-1}E$, $\beta=I^TD^{-1}E$, $\gamma=I^TD^{-1}I$ for short. The received equation system gets the following form

$$2\theta+\lambda\alpha+\mu\beta=0,$$

$$2+\lambda\beta+\mu\gamma=0.$$

From these equations a dependence between c and θ can be derived

$$c_k=\frac{\theta\gamma-\beta}{\alpha\gamma-\beta^2}\sum_{i=1}^{L}d_{ki}V\left(z_i\right)+\frac{\theta\beta-\alpha}{\beta^2-\alpha\gamma}\sum_{i=1}^{L}d_{ki}>0,\ k=1,...,L, \tag{3}$$

where d_{ij} is an element of the D^{-1} matrix.

It must be noted also that the point c^* can be a point of strict maximum of P_f only if

$$\sum_{i=1}^{L}\sum_{j=1}^{L}\rho_{ij}\varepsilon_i\varepsilon_j>0 \tag{4}$$

for any $\left(\varepsilon_0,...,\varepsilon_L\right)$ satisfying conditions $\sum_{i=1}^{L}\varepsilon_i=0$. Let $\theta_{\min}$ is minimal and $\theta_{\max}$ is maximal value of θ for which one of inequalities (3) becomes equality. Let $R_k^V=\sum_{i=1}^{L}V\left(z_i\right)\rho_{ki}$, $P_k=\sum_{i=1}^{L}\rho_{ki}$,

$$\Gamma_i^1 = \frac{\gamma R_i^v + \beta P_k}{\alpha\gamma - \beta^2},$$

$$\Gamma_i^0 = \frac{\alpha R_i^v + \beta P_k}{-\alpha\gamma + \beta^2}.$$

then $P_f = B_0 + B_1\theta + B_2\theta^2$, where

$$B_0 = \sum_{i=1}^{L}\sum_{j=1}^{L}\Gamma_i^0\Gamma_j^0\rho_{ij},$$

$$B_1 = \sum_{i=1}^{L}\sum_{j=1}^{L}\left(\Gamma_i^0\Gamma_j^1 + \Gamma_i^1\Gamma_j^0\right)\rho_{ij},$$

$$B_2 = \sum_{i=1}^{L}\sum_{j=1}^{L}\Gamma_i^1\Gamma_j^1\rho_{ij}.$$

It is easy to show that

$$K\left(Y, Z_{ccp}\right) = \kappa(\theta) = \frac{1}{\sqrt{V(Y)}}\frac{\theta}{\sqrt{B_1\theta - B_2\theta^2 - B_0}}.$$

Theorem 2. Simultaneous correctness of inequalities $\theta_{\min} < \frac{2B_0}{B_1} < \theta_{\max}$, $\kappa\left(\frac{2B_0}{B_1}\right) > \kappa\left(\theta_{\min}\right)$ and negativity of the condition (4) is necessary condition of irreducibility of predictors set $z_1,\ldots,z_L$.

Necessary conditions allows effectively evaluate irreducibility of predictors set. It is sufficient to calculate $\theta_{\min}$ and $\theta_{\max}$ to evaluate negativity conditions (4) and to evaluate inequalities $\theta_{\min} < \frac{2B_0}{B_1} < \theta_{\max}$. It is evident that in case when necessary conditions are satisfied and $\kappa\left(\frac{2B_0}{B_1}\right)$ for the evaluated ensemble is greater than maximal correlation coefficient for any irreducible ensemble with less predictors than the evaluated ensemble is irreducible. It is important that optimal coefficients c_k may be received from (3) when $\theta = \frac{2B_0}{B_1}$.

Regression models based on sets of unexpandable irreducible ensembles

At the first stage initial set of reduced predictors is formed with the help of standard uni-variate least squares technique. Let $\tilde{Z} = \left(z_1,\ldots,z_L\right)$ is initial set of L predictors. An irreducible ensemble $\tilde{z}'$ consisting of l' predictors will be called unexpandable irreducible ensemble (UIE) if there are no irreducible ensembles in $\tilde{Z}$ with number of predictors greater l' that contain all predictors from $\tilde{z}'$. Two ways of regression model construction by sets of UIE were considered that are based on enumerating of all possible UIE. The first method chooses single best UIE where correlation coefficient of optimal Z_{ccp} with Y is maximal. This optimal Z_{ccp} ($Z_{ccp}^{\max}$) is the final regression model of the first method. The second method selects set of UIE where correlation coefficient of optimal Z_{ccp} with Y is greater than $Tr * K\left(Y, Z_{ccp}^{\max}\right)$, $Tr \in (0,1)$. Thus threshold parameter Tr allows to select UIE with correlation coefficient of optimal Z_{ccp} with Y close to maximal value $K\left(Y, Z_{ccp}^{\max}\right)$. It is supposed that Tr must be close to 1. In the second method parameters of final regression models are calculated as average by all UIE with $K\left(Y, Z_{ccp}\right) > Tr * K\left(Y, Z_{ccp}^{\max}\right)$. Our experiments showed that second approach is more effective. Method of UIE enumerating is based on gradual raising of predicates set meeting irreducibility condition. First, a set of all possible predictor pairs P_2 is considered. A set of all irreducible pairs P_2^{irr} is then extracted using the Theorem 2 results. Subsequently, a set of triplets P_3^{irr} is formed using P_2^{irr}. The process is going on until step i in which P_i^{irr} becomes empty. UIE based method depends on squared variances of single predictors V and distances between predictors ρ. The parameters were evaluated from training data by standard formulae $V(z) = \frac{1}{M}\sum_{j=1}^{M}\left[z_j - \hat{z}\right]^2$, $\rho\left(z^1, z^2\right) = \frac{1}{M}\sum_{j=1}^{M}\left[z_j^1 - z_j^2\right]^2$, where M is training set size. Experiments showed that such type of estimates leads to selection of too many variables and so to decrease of prognostic ability. However effectiveness may be systematically improved by using additional penalty multiplier for ρ equal $\frac{1}{1+\frac{5}{M}}$. This effect demands mathematical explanation.

Experiments

In all studies dependent variable Y and regressor variables X are stochastic functions of 3 latent variables U_1, U_2, U_3. The vector levels of variables U are independently distributed multivariate normal with mean 0 and standard deviation 1. The value of dependent variable Y in j-th case is generated by formula $y_j = \sum_{k=1}^{3} u_{jk} + e_y^j$ where u_{jk} is a value of the latent variable U_k, e_y^j is a random error term distributed $N\left(0, d_y\right)$. At that 85% of cases were generated with $d_y = 1$, 15% of cases were generated with $d_y = 2$. Thus, main and noisy components of data were modelled. The values of relevant variable X_i were generated by binary vector $\beta^i = \left\{\beta_1^i, \beta_2^i, \beta_3^i\right\}$. In j-th case $x_{jk} = \sum_{k=1}^{3} u_{jk}\beta_k^i + e_{xi}^j$, where u_{jk} is a value of the latent variable U_k, e_{xi}^j is a random error term distributed $N\left(0, d_{xi}\right)$. At that for 5 relevant variables $d_{xi} = 0.2$ and rest relevant variables were generated according $d_{xi} = 0.5$. The levels of irrelevant variable X_i in j-th case are generated by formula $x_{jk} = e_{xi}^j$. In each experiment 100 pairs of data sets were calculated by the random numbers generator according to the same scenario. Each pair includes training set that was used for optimal regression model construction and control data set that was used to evaluate prognostic ability of this model. In all experiments relevant variables were generated at $\beta = \{1,1,0\}$, $\beta = \{1,0,1\}$, $\beta = \{0,1,1\}$. Results of experiments are given in the Table 1. For each pair of samples of size M the following characteristics of forecasting ability for LARS and multiple UIE regression with $Tr = 0.95$ are given: K – correlation coefficient between variable Y and calculated prognoses, N_c – average number of relevant variables that were correctly used in regression model, N_f – average number of irrelevant variables that were mistakenly used in regression model, R_f – ratio of $\frac{|\beta|}{|\beta_{\max}|}$ for irrelevant variables. Here $|\beta|$ is an absolute value of regression coefficient for some variable in regression model, $|\beta_{\max}|$ is the maximal absolute value of regression coefficient among variables of regression model. It was considered that variable v_i is used by regression model if corresponding ratio $\frac{|\beta|}{|\beta_{\max}|}$ is less than 0.001.

Table 1. Results of expiriments. Prognostic ability.

	CCP_{cor}				CCP_{error}			
M	K	N_c	N_f	R_f	K	N_c	N_f	R_f
20	0.729	15.13	6.03	0.132	0.694	4.84	0.23	0.092
30	0.752	16.81	5.76	0.054	0.716	6.57	0.07	0.006
40	0.772	17.36	7.35	0.066	0.744	8.34	0.03	0.012
50	0.776	17.21	5.69	0.03	0.742	9.27	0	0

Conclusion

The results shown in the table 1 clearly show the superiority of the described novel approach, i.e. correction based on correlation maximization, over previously described [Senko et al., 2010] error minimization based one. Namely the correlation K is about 0.03 higher in all tasks. It is achieved by correct selection of almost all informative variables of the samples and though the amount of falsely selected noise variables is also increased, their weights in resulting combinations are low.

The primary drawback of the proposed method is slow speed that is decreasing dramatically with the increase of a task dimension. It is planned that further research will be aimed at reduction of computational complexity it. Nevertheless, the method proved to be suitable for a wide range of forecasting applications, especially in tasks which require feature selection.

Bibliography

[Efron et al., 2004] B. Efron, T. Hastie, I. Jonnstone and R. Tibshirani. Least Angle Regression. Annals of Statistics. 2004, Vol. 32, No. 2, 407–499.

[Tibshirani, 1996] Tibshirani R., Regression shrinkage and selection via the lasso // J. Roy. Stat. Soc. 1996. Vol. 58, p. 267–288.

[Breiman, 1999] L. Breiman, Random forests - random features. Technical report 567. Statistics department. University of California, Berkley, September 1999 // www.boosting.org.

[Kuncheva, 2004] L.I. Kuncheva, Combining Pattern Classifiers. Methods and Algorithms. Wiley Interscience, New Jersey, 2004.

[Zhuravlev et al., 2008] Zhuravlev Yu.I., Kuznetsova A.V., Ryazanov V.V., Senko O.V., Botvin M.A., The Use of Pattern Recognition Methods in Tasks of Biomedical Diagnostics and Forecasting // Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. 2008, Vol. 18, No. 2, pp. 195–200.

[Zhuravlev et al., 2006] Zhuravlev Yi.I., Ryazanov V.V., Senko O.V., RECOGNITION. Mathematical methods. Program System. Applications. —Moscow: Phasiz, 2006, (in Russian).

[Kuznetsov et al., 1996] Kuznetsov V.A., Senko O.V. et all., Recognition of fuzzy systems by method of statistically weighed syndromes and its using for immunological and hematological norm and chronic pathology // Chemical Physics, 1996, v. 15, N 1, p. 81–100.

[Brown et al., 2005] Gavin Brown, Jeremy L. Wyatt, Peter Tino, Managing Diversity in Regression Ensembles. Journal of Machine Learning Research 6: 1621-1650. 2005.

[Krogh et al., 1995] A. Krogh and J. Vedelsby, Neural network ensembles, cross validation, and active learning. NIPS, 7:231–238, 1995.

[Senko, 2004] Senko Oleg V., The Use of Collective Method for Improvement of Regression Modeling Stability // InterStat. Statistics on the Internet http://statjournals.net/, June, 2004.

[Senko, 2009] O.V. Senko,. An Optimal Ensemble of Predictors in Convex Correcting Procedures // Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. 2009, Vol. 19, No. 3, pp. 465–468.

[Senko et al., 2010] Senko O., Dokukin A. Optimal Forecasting Based on Convex Correcting Procedures.// New Trends in Classification and Data Mining -ITHEA, Sofia, Bulgaria, 2010, p. 62-72.

Acknowledgements

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Authors' Information

Oleg Senko – CCAS, chief researcher, 119333, Vavilova Str. 40, Moscow, Russian Federation; e-mail: senkoov@mail.ru

Major Fields of Scientific Research: Pattern Recognition, Data Mining

Alexander Dokukin – CCAS, researcher, 119333, Vavilova Str. 40, Moscow, Russian Federation; e-mail: dalex@ccas.ru

Major Fields of Scientific Research: Pattern Recognition, Data Mining