ON SOME PROPERTIES OF REGRESSION MODELS BASED ON CORRELATION MAXIMIZATION OF CONVEX COMBINATIONS

Oleg Senko, Alexander Dokukin

Abstract: The article is devoted to thorough study of a new regression method performance. The proposed method based on convex correcting procedures over sets of predictors is subject to modifications and tested in comparison with the acknowledged regression utility. The modifications touch both resource consumption and quality aspects of the method and tests are performed with sets of generated samples.

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

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

Introduction

Present article continues a series of works devoted to an approach in which optimal forecasting models are built by large ensembles of preliminary trained predictors that in turn can be simple univariate egressions. 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 [Tibshirani, 1996], [Efron et al., 2004] methods may be mentioned thereupon. However we believe that a problem of low generalization ability of empirical models in high-dimensional tasks cannot be considered completely solved. Thus, a number of convex correcting procedures optimization method has been proposed [Senko, 2009], [Senko et al., 2010], [Senko et al., 2011].

Suppose that we have set of L predictors $z_1, \ldots, z_L$ that forecast some variable Y. Let $c = (c_1, \ldots, c_L)$ 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].

A method for CCP optimization that is based on minimization of general error estimates was studied in [Senko, 2009], [Senko et al., 2010]. Experiments with simulated data demonstrated that CCP error optimization also implements effective selection of informative prognostic variables.

In [Senko et al., 2011], however, it was shown that CCP variance is decreased comparing to the same combination of single predictors' variances and such a decrease deteriorates the CCP's prognostic ability. An additional adjustment to be made to CCP predictions leads to the necessity of maximizing Z_{ccp} and Y correlation. Such a technique based on the same concept of irreducible ensembles searching that was used in [Senko et al., 2010] was proposed in the article.

Again, experiments with simulated data demonstrated that CCP correlation optimization shows great results comparing to LARS method, the only drawback of the result being that LARS was implemented by the authors and thus may be not the optimal one. So, in present article the method is compared to widely acknowledged Glmnet for Matlab written by Jerome Friedman and Hui Jiang [Friedman et al., 2007], [Friedman et al., 2010].

In the next few sections we afford repeating some definitions and theorems concerning irreducible ensembles searching and convex correctors' correlation optimization. Then, some modifications to the correlation maximization method (CCPCMM) will be described. And finally, the results of experiments will be shown.

Irreducible ensembles relative to correlation coefficients

It is supposed further that predictors from initial set are additionally transformed with the help of optimal univariate 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 $\hat{X}$. It is known that following inequalities are true for a reduced predictor z:

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

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 univariate correcting transformation becomes inevitable. Of all predictors the maximal quality is provided by the one most correlated with Y.

Standard Pearson correlation coefficient is defined as the ratio:

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

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

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

where ρ_{ij} denotes discrepancy between i-th and j-th predictors.

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

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

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

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

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

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

$$P_f(c) = \sum_{i=1}^{L} \sum_{j=1}^{L} c_i c_j \rho_{ij}^v.$$

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

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

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

with the next contingencies:

$$\sum_{i=1}^{L} c_i E(z_i^2) = \theta,$$

$$\sum_{i=1}^{L} c_i = 1, \tag{2}$$

$$c_i \geq 0, \ i = 1, \ldots, L.$$

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

$$L = \sum_{i=1}^{L} \sum_{j=1}^{L} c_i c_j \rho(z_i, z_j) + \lambda \left(\sum_{i=1}^{L} c_i E(z_i^2) - \theta \right) + \mu \left(\sum_{i=1}^{L} c_i - 1 \right),$$

and equal its partial derivatives to zero

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

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{L} c_i E\left(z_i^2\right) - \theta = 0 ,$$

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

Moving to a vectorial form we get

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

$$E^T C = \theta ,$$

$$I^T C = 1 .$$

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

Lets denote $\alpha = E^T D^{-1} E$, $\beta = I^T D^{-1} E$, $\gamma = I^T D^{-1} I$ 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

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

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 $R_k = \sum_{i=1}^{L} d_{ki} E\left(z_i^2\right)$, $P_k = \sum_{i=1}^{L} \rho_{ki}$,

$$c_k = \theta\Gamma_k^1 + \Gamma_k^0$$

$$\Gamma_k^1 = \frac{\gamma R_k - \beta P_k}{\alpha\gamma - \beta^2} ,$$

$$\Gamma_k^0 = \frac{\alpha P_k - \beta R_k}{\alpha\gamma - \beta^2} .$$

then $P_f = \beta_0 + \beta_1\theta + \beta_2\theta^2$, where

$$\beta_0 = \sum_{i=1}^{L} \sum_{j=1}^{L} \Gamma_i^0 \Gamma_j^0 \rho_{ij} ,$$

$$\beta_1 = \sum_{i=1}^{L}\sum_{j=1}^{L}\left(\Gamma_i^0\Gamma_j^1 + \Gamma_i^1\Gamma_j^0\right)\rho_{ij},$$

$$\beta_2 = \sum_{i=1}^{L}\sum_{j=1}^{L}\Gamma_i^1\Gamma_j^1\rho_{ij}.$$

It is easy to show that

$$K\left(Y, Z_{ccp}\right) = \kappa(\theta) = \frac{1}{\sqrt{V(Y)}}\frac{\theta}{\sqrt{\beta_0 + \beta_1\theta + \beta_2\theta^2}}.$$

Theorem 2. Simultaneous correctness of inequalities $\theta_{\min} < \dfrac{2\beta_0}{1-\beta_1} < \theta_{\max}$, $\kappa\left(\dfrac{2\beta_0}{1-\beta_1}\right) > \kappa(\theta_{\min})$ 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 $\theta_{\min} < \dfrac{2\beta_0}{1-\beta_1} < \theta_{\max}$. It is evident that

in case when necessary conditions are satisfied and $\kappa\left(\dfrac{2\beta_0}{1-\beta_1}\right)$ for the evaluated ensemble is greater than

maximal correlation coefficient for any irreducible ensemble with less predictors than the evaluated ensemble is
irreducible.

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 univariate 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 $(1 - Tr)K\left(Y, Z_{ccp}^{\max}\right)$, $Tr \in (0,1)$. Thus threshold parameter Tr allows to

select UIE with correlation coefficient of optimal Z_{ccp} with Y close to maximal value $K\left(Y, Z_{ccp}^{\max}\right)$. In the
second method parameters of final regression models are calculated as average by all UIE with
$K\left(Y, Z_{ccp}\right) > Tr * K\left(Y, Z_{ccp}^{\max}\right)$.

Method of UIE enumerating is based on gradual raising of predicates set meeting irreducibility condition.

Procedure 1. Process subset of predictors $Z = \left(z_{i_1}, \ldots, z_{i_t}\right)$.

Step 1. Using Theorem 2 check whether Z is irreducible.

Step 2. Calculate $\left(c_{i_1}, \ldots, c_{i_t}\right)$ and $K(Y, Z)$.

Step 3. If $K(Y,Z) > K^*$, where K^* is the previous best result, replace best subset Z_{ccp}^{max} with Z and set $K^* = K(Y,Z)$.

Step 4 (second method only). Store Z in historic list for voting purposes.

Procedure 2. Main algorithm.

Step 1. Enumerate all pairs of predictors (z_i, z_j), apply Procedure 1. If (z_i, z_j) is irreducible, store it in pairs dictionary and in list of candidates.

Step 2. Enumerate all current candidates $Z = (z_{i_1}, \ldots, z_{i_t})$, enumerate all pairs from dictionary, beginning with $z_{i_t} : (z_{i_t}, z_k)$. Apply Procedure 1 to the subset $Z' = (z_{i_1}, \ldots, z_{i_t}, z_k)$. If it is irreducible, store it in next level candidates.

Step 3. If there are any next level candidates, go to Step 2. Otherwise stop and return current Z_{ccp}^{max} (and historic list).

Step 4 (second method only). Filter historic list based on K^* and Tr and average coefficients over all remaining combinations. Let's consider a set of combinations produced by the algorithm: $\{Z_1, \ldots, Z_p\}$, where

$$Z_t = \sum_{i=1}^{L} c_i^t z_i .$$ The final predictor $Z = \sum_{i=1}^{L} \frac{1}{p} \left(\sum_{j=1}^{p} c_i^j \right) z_i .$

CPPCMM modifications

First of all, lets state that only second method, i.e. voting over some set of best combinations, is considered as proved to be better in experiments.

A new set of experiments performed for the sake of this article has revealed a major drawback of the described method. Significant time consuming in cases of larger dimensions was accompanied by memory exhaustion. Thus, strict UIE enumerating demanded additional branch reducing:

Procedure 3. Reduced main algorithm.

Step 1. Enumerate all pairs of predictors (z_i, z_j), apply Procedure 1. If (z_i, z_j) is irreducible, store it in pairs dictionary and in list of candidates.

Step 2. Consider level l. Enumerate all current candidates $Z = (z_{i_1}, \ldots, z_{i_t})$, enumerate all pairs from dictionary, beginning with $z_{i_t} : (z_{i_t}, z_k)$. Apply Procedure 1 to the subset $Z' = (z_{i_1}, \ldots, z_{i_t}, z_k)$. If $K(Y,Z') > K_{l+1}^*$, where K_{l+1}^* is the previous best result of $l+1$ level, set $K_{l+1}^* = K(Y,Z')$. If Z' is irreducible and $K(Y,Z') > K_l^*$, store it in $l+1$ level candidates.

Step 3. If there are any next level candidates, go to Step 2. Otherwise stop and return current Z_{ccp}^{max} (and historic list).

Step 4. Filter historic list based on K^* and Tr and average coefficients over all remaining combinations.

The proposed correction although provided giant boost in time and memory saving, slightly dropped overall forecasting quality. The next two modifications are aimed to its correction.

Definition 2. A predictor z_i is dominating z_j if $K(Y,z_i) \geq aK(Y,z_i) + bK(Y,z_j)$ for all $a,b > 0$, $a + b = 1$.

Theorem 3. A predictor z_i is dominating z_j if $\dfrac{V(z_j)^2 - V(z_i)V(z_j) - V(z_j)\rho_{ij}}{\left(V(z_i) - V(z_j)\right)^2 - \left(V(z_i) + V(z_j)\right)\rho_{ij}} > 1$.

The second modification consists in removing all dominated predictors from voting according to Theorem 3.

Third modification is weighting votes of different predictors in Step 4: $Z = \sum\limits_{i=1}^{L}\left(\sum\limits_{j=1}^{p} w_j c_i^j\right) z_i$, where $w_k \geq 0$ and

$\sum\limits_{k=1}^{p} w_k = 1$.

In case of no domination filter applied the weights are calculated simply in proportion to $\dfrac{1}{\left(1 - K(Y,Z_i)^2\right)}$.

Otherwise they are more complicated. Let $w_{ij} = \dfrac{\left(\rho_{ij} + e_j - e_i\right)}{2\rho_{ij}p}$ and consequently $w_i = \sum\limits_{j=1}^{p} w_{ij}$. Again,

normalization is applied to satisfy $\sum\limits_{k=1}^{p} w_k = 1$ condition.

With that last modification, the parameter Tr (threshold) described in previous section, although planned as close to zero, proved to be more efficient when close to 1 (see experiments).

Experiments

In all studies dependent variable Y and regression 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 $y_j = \sum\limits_{k=1}^{3} u_{jk} + e_y^j$ where u_{jk} is a value of the latent variable U_k, e_y^j is a random error term distributed $N(0, d_y)$. At that 85% of cases were generated with $d_y = 1$, 15% of cases were generated with $d_y = 2$ or $d_y = 2.5$. That is how main and noisy components of data were formed. 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 $x_{jk} = \sum\limits_{k=1}^{3} u_{jk}\beta_k^i + e_{xi}^j$, where u_{jk} is a value of the latent variable U_k, e_{xi}^j is a random error term distributed $N(0, d_{xi})$. In the following experiments 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 $\beta = \{1,1,0\}$, $\beta = \{1,0,1\}$, $\beta = \{0,1,1\}$. In Table 1 there are other parameters of the test samples described.

Table 1. Experiment sample series.

Task	Number of objects	Number of features	...of them irrelevant	Noize coefficient d_y
data1	30	120	70	2.0
data2	30	120	70	2.5
data3	30	100	50	2.5
data4	30	140	90	2.0
data6	20	160	85	2.0
data7	20	160	85	2.5
data8	15	150	85	2.1
data9	40	150	81	2.5

First, the described data was used for the threshold parameter Tr impact study. The following two graphs show the dependency between resulting forecast correlation and the parameter. Here and further on an average values over 100 independent control tasks are shown.
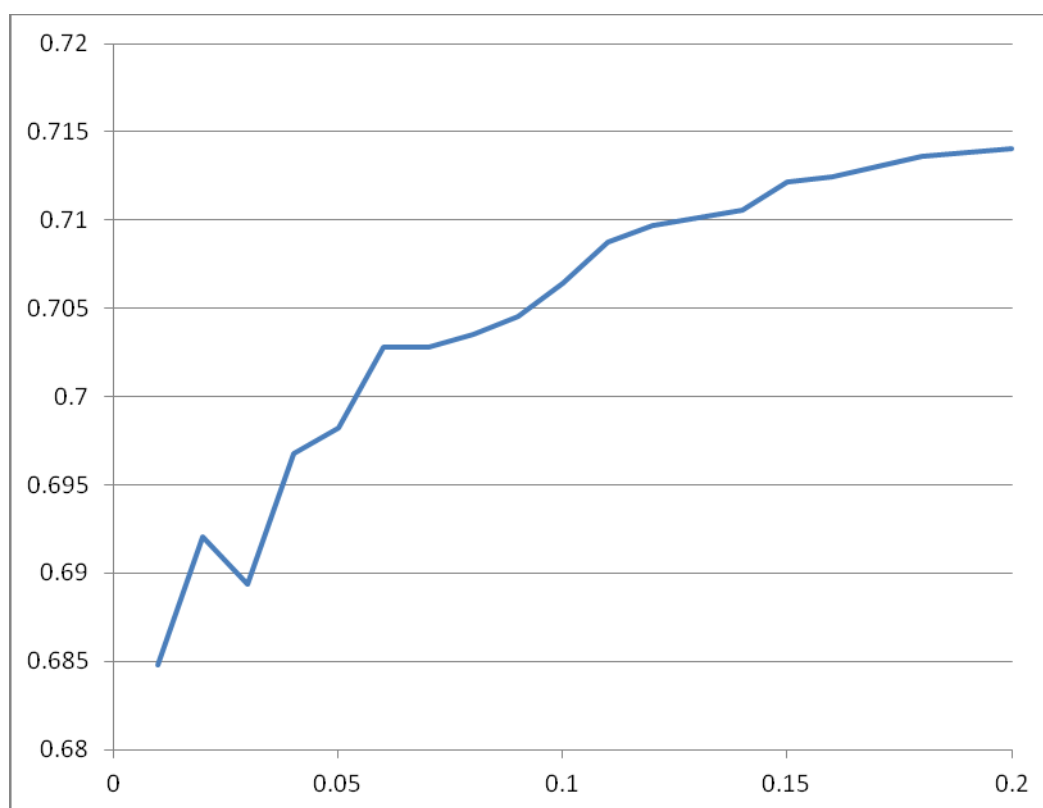
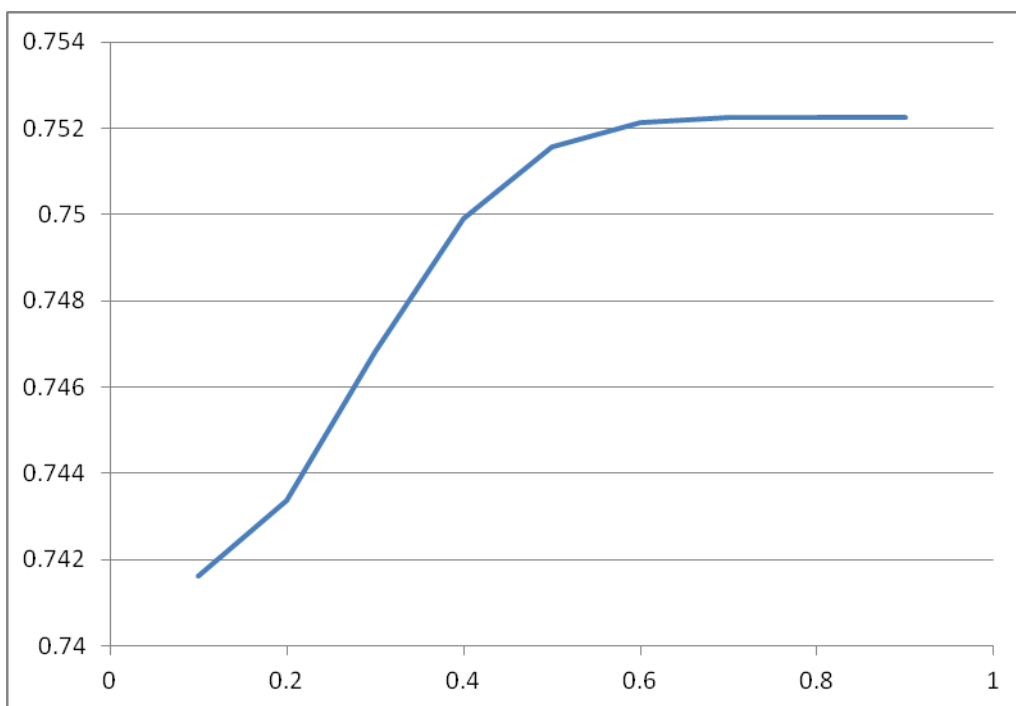**Fig. 1. Data7 test sample, threshold range 0.1–0.2, step 0.01.**

Fig. 2. Data8 test sample, threshold range 0.1–0.9, step 0.1.

The clear and most unexpected result of this and other similar experiments is that better result are achieved at threshold values close to 1. It means that every tested irreducible combination is important for the resulting weighted sum. Furthermore, the dependency is quite monotonic and thus threshold in every comparative experiment can be set to 1.

Finally, the second set of experiments shows comparison of the proposed method to Glmnet. It need to be mentioned that Glmnet for Matlab also has some parameters. Thus, to make results more undoubted its optimization was performed, so all tables and graphs contain its best result over parameters grid.

Table 2. Results of expiriments. Prognostic ability.

Task	CCPCMM's correlation	Glmnet for Matlab	
		Correlation	Optimal parameter λ
data1	0.776	0.763	0.55
data2	0.746	0.726	0.5
data3	0.741	0.722	0.55
data4	0.752	0.739	0.55
data6	0.768	0.736	0.55
data7	0.728	0.691	0.75
data8	0.752	0.7135	0.75
data9	0.732	0.711	0.57

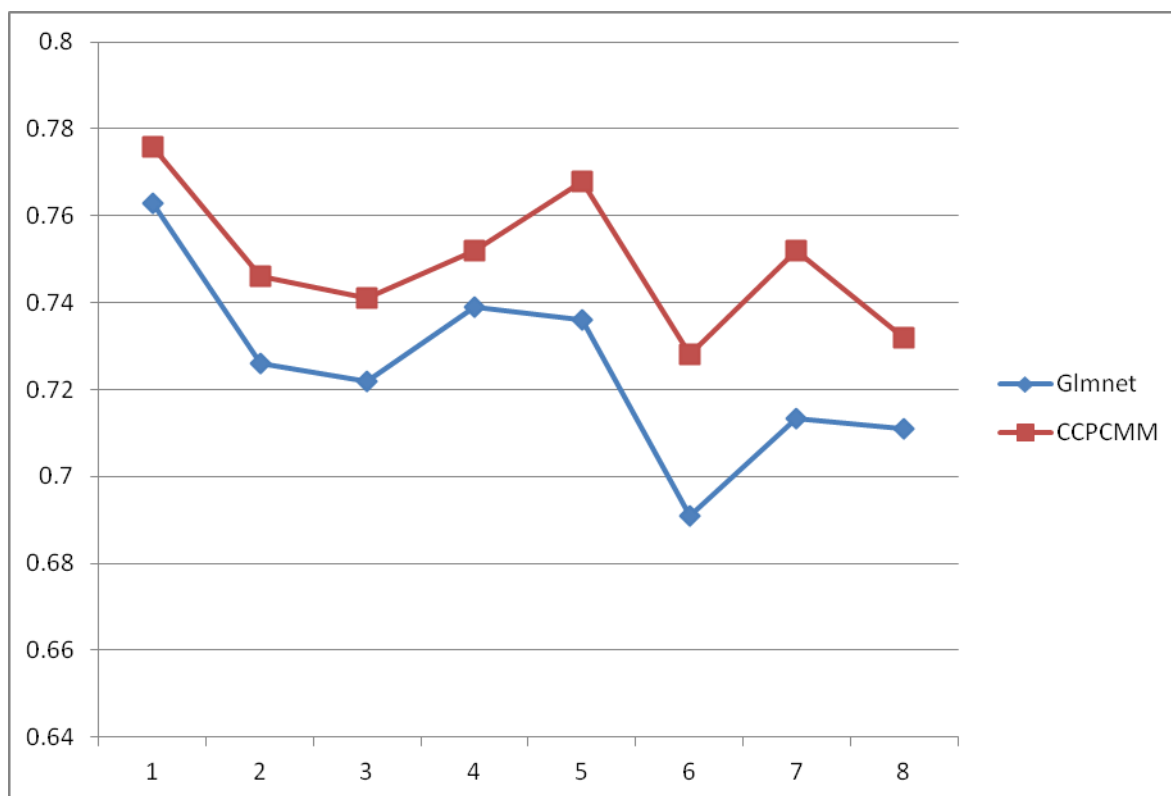The same results are shown on the following graph.

Fig. 3. Correlation in test tasks.

Conclusion

Some modifications of the novel regression method are described, which correct its time and memory consuming as well as forecasting quality. The results shown in figures 1 and 2 exclude any parameters from the training process, which made it suitable for unsupervised use. Moreover, the results shown in table 2 and figure 3 clearly show its superiority comparing to well known and widely acknowledged regression tool.

Thus, the modified method can be recommended for a wide range of forecasting applications, especially in automatic unsupervised applications.

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.

[Senko et al., 2011] Senko O., Dokukin A. Correlation Maximization in Regression Models Based on Convex Combinations // International Journal "Information Theories and Applications", Vol. 18, Number 3, 2011, P. 224-231.

[Friedman et al., 2007] Jerome Friedman, Trevor Hastie, Holger Hofling, Robert Tibshirani. Pathwise coordinate optimization, The Annals of Applied Statistics. Volume 1, Number 2 (2007), 302-332.

[Friedman et al., 2010] Jerome Friedman, Trevor Hastie, Rob Tibshirani. Regularized Paths for Generalized Linear Models via Coordinate Descent // Journal of Statistical Software 33(1), 2010.

Acknowledgements

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