Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

New Trends
in
Classification and Data Mining

ITHEA
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0042-9

C\o Jusautor, Sofia, 2010

OPTIMAL FORECASTING BASED ON CONVEXCORRECTING PROCEDURES

Oleg Senko, Alexander Dokukin

***Abstract**: Properties of convex correcting procedures (CCP) on ensembles of predictors are studied. CCP calculates integral solution as convex linear combination of predictors' prognoses. Structure of forecasting squared error and generalized error are analyzed. At that generalized error is defined as mean of squared error at Cartezian product of forecasted objects space and space of training sets. It is shown that forecasting squared error, bias and variance component of generalized error have similar structure. Search of optimal CCP coefficients is reduced to quadratic programming task which is solved in terms of ensemble superfluity. Ensemble is considered superfluous if some members can be removed without loss of forecasting ability. Necessary and sufficient conditions of superfluity absence are proven. A regression method based on the described principles has been developed. Its concepts as well as testing results are shown revealing CCP's significant superiority over stepwise regression.*

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

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

Introduction

Goal of this work is study of correcting procedures for sets of forecasting algorithms calculating integral solution as convex linear combination of prognoses, given by each algorithm from the set. Let's suppose that we have set of L algorithms, forecasting some variable Y by set of explanatory variables $X_1,....,X_n$ at objects that are elements of probability space Ω. Prognosis that is calculated by *i-th* algorithm for some ω will be further denoted as $z_i(\omega)$. Let $\mathbf{c}=(c_1,\ldots,c_L)$ is vector of real nonnegative coefficients satisfying $\sum_{i=1}^{L} c_i = 1$. Convex correcting procedures (CCP) are discussed in the paper that calculate collective solution $Z(\omega,\mathbf{c})$ as $Z_{ccp}(\omega,\mathbf{c})=\sum_{i=1}^{L} c_i z_i(\omega)$. Using of average by set of prognoses is special case of CCP.

Convex correcting procedures are rather often used in theory of pattern recognition or forecasting by empirical data. Neural networks ensembles, methods of weighed combining, boosting and bagging methods, pattern recognition methods based on voting by systems of regularities [1, 2] are well known examples of optimal convex solutions. Last years some new techniques were suggested that are based on searching balance between accuracy of data approximation and diversity of ensembles [5]. Our approach is based on analysis of generalized error structure.

Forecasting error for CCP

We begin with discussing of mean squared error of CCP forecasting. Mean squared error of Y forecasted by some predictor Z will de denoted as $\delta(Z)$. So, $\delta(Z)=E_{\Omega}(Y-Z)^2$. It is easily to show that

$$\sum c_i[Y(\omega) - z_i\,(\omega)]^2 = \sum_{i=1}^{l} c_i[Y(\omega) - Z_{ccp}(\omega,\mathbf{c}) + Z_{ccp}(\omega,\mathbf{c}) - z_i\,(\omega)]^2 =$$

$$= \sum_{i=1}^{L} c_i[Y(\omega) - Z_{ccp}(\omega,\mathbf{c})]^2 + \sum_{i=1}^{L} c_i[Z_{ccp}(\omega,\mathbf{c}) - z_i(\omega)]^2 =$$

$$= [Y(\omega) - Z(\omega,\mathbf{c})]^2 + \sum_{i=1}^{L} c_i[Z_{ccp}(\omega,\mathbf{c}) - z_i(\omega)]^2$$

So CCP error in case of forecasting of Y for object ω that is equal $[Y(\omega) - Z_{ccp}(\omega,\mathbf{c})]^2$ can be presented as difference

$$\sum_{i=1}^{L} c_i[Y(\omega) - z_i(\omega)]^2 - \sum_{i=1}^{L} c_i[Z_{ccp}(\omega,\mathbf{c}) - z_i\,]^2 \qquad (1)$$

Decomposition (1) was received in [4].

Task of optimal CCP search may be discussed as task of minimization of mathematical mean of error $[Y(\omega) - Z(\omega,\mathbf{c})]^2$ in space of forecasted objects. It is evident that

$$\delta(Z_{ccp}) = E_\Omega\{\sum_{i=1}^{L} c_i[Y(\omega) - z_i(\omega)]^2 - \sum_{i=1}^{L} c_i[Z_{ccp}(\omega,\mathbf{c}) - z_i\,]^2\} = \sum_{i=1}^{L} c_i\delta(z_i) -$$

$$-\sum_{i=1}^{L} c_i[Z_{ccp}(\omega,\mathbf{c}) - z_i\,]^2\}.$$

It follows from non-negativeness of variation component $E_\Omega\{\sum_{i=1}^{L} c_i[Z_{ccp}(\omega,\mathbf{c}) - z_i(\omega)]^2\}$ that error $\delta(Z_{ccp})$ never exceed weighed with $c_i > 0$ mean of individual prognostic algorithms errors. Mathematical mean $E_\Omega[z_i - z_j]^2$, characterizing discrepancy of i'-th and j-th forecasting algorithms will be denoted as $E_\Omega[z_i - z_j]^2 = \rho_{ij}^e$. Let note that

$$-\sum_{i=1}^{L} c_i[Z_{ccp}(\omega,\mathbf{c}) - z_i\,]^2 = E_\Omega[Z(\omega,\mathbf{c})]^2 - \sum_{i=1}^{L} c_i E_\Omega[z_i^2(\omega)] =$$

$$= \sum_{i'=1}^{L}\sum_{i''=1}^{L} c_{i'}c_{i''}E_\Omega[z_{i'}(\omega)z_{i''}(\omega)] - \sum_{i=1}^{L} c_i E_\Omega[z_i^2(\omega)].$$

Taking into account that $z_{i'}z_{i''} = \frac{1}{2}\{-(z_{i'} - z_{i''})^2 + (z_{i'})^2 + (z_{i''})^2\}$, we receive that

$$\sum_{i=1}^{L}\sum_{i=1}^{L} c_{i'}c_{i''}E_\Omega[z_{i'}(\omega)z_{i''}(\omega)] - \sum_{i=1}^{L} c_i E_\Omega[z_i^2(\omega)] = -\tfrac{1}{2}\sum_{i=1}^{L}\sum_{i=1}^{L} c_{i'}c_{i''}E_\Omega[z_{i'}(\omega) - z_{i''}(\omega)]^2 +$$

$$+\tfrac{1}{2}\{\sum_{i=1}^{L} c_{i'}E_\Omega[z_{i'}(\omega)]^2 \sum_{i=1}^{L} c_{i''} + \sum_{i''=1}^{L} c_{i''}E_\Omega[z_{i''}(\omega)]^2 \sum_{i'=1}^{L} c_{i'}\} - \sum_{i=1}^{L} c_i E_\Omega[z_i^2(\omega)] =$$

$$+\sum_{i''=1}^{L} c_{i''} E_{\Omega}[z_{i''}(\omega)]^2 \sum_{i'=1}^{L} c_{i'}\} - \sum_{i=1}^{L} c_i E_{\Omega}[z_i^2(\omega)] = -\sum_{i'=1}^{L}\sum_{i''=1}^{L} c_{i'} c_{i''} E_{\Omega}[z_{i'} - z_{i''}]^2 =$$

$$= -\sum_{i'=1}^{L}\sum_{i''=1}^{L} c_{i'} c_{i''} \rho^{e}_{i'i''} \tag{2}$$

So, error of CCP forecasting may be written as

$$\delta(Z_{ccp}) = \sum_{i=1}^{L} c_i \delta(z_i) - \tfrac{1}{2}\sum_{i'=1}^{L}\sum_{i''=1}^{L} c_{i'} c_{i''} \rho^{e}_{i'i''} \tag{3}$$

Generalized forecasting error of CCP

Generalized error. Forecasting error $\delta(Z)$ describes exactness of forecasting algorithm (predictor) Z that was previously trained by some fixed training set ω_t. But training set often may change during process of training. In these cases predictor Z is function of ω and ω_t. Effectiveness of training procedure is better characterized with help of generalized error that is mathematical mean of error $\delta(Z)$ by space of various training sets Ω_t. The generalized error for predictor Z will be denoted as $\Delta(Z)$. The mean value Z at point of $\mathbf{x} \in \mathbf{R}^n$ by space Ω_t will be denoted as $\hat{Z}(\mathbf{x})$. The following decomposition is true for generalized error $\Delta(Z)$:

$$\Delta(Z) = \Delta_{noise} + \Delta_{bias}(Z) + \Delta_{\mathrm{var}}(Z),$$

where $\Delta_{noise} = E_{\Omega}\{Y - E[Y \mid \mathbf{x}(\omega)]\}^2$ is irreducible noise component that characterize only random process associated with each specific forecasting task and is not related to forecasting algorithm, component $\Delta_{bias}(Z) = E_{\Omega}\{\widehat{Z}[\mathbf{x}(\omega)] - E_{\Omega}[Y \mid \mathbf{x}(\omega)]\}^2$ describes deviation of $\hat{Z}(\mathbf{x})$ from conditional means $E_{\Omega}[Y \mid \mathbf{x}(\omega)]$, $\Delta_{\mathrm{var}} = E_{\Omega} E_{\Omega_t}\{\widehat{Z}[\mathbf{x}(\omega)] - Z[\mathbf{x}(\omega, \omega_t)]\}^2$ describes variation of $Z[\mathbf{x}(\omega, \omega_t)]$ at Cartesian product $\Omega_t \times \Omega$. At that bias component is related to inconsistency between type of used model and dependency that really exists in training data set, while variance component is related to inconsistency between complexity and dimension of model and training data set size. Variance component describes variation of forecasting function at relatively small and statistically admissible changes in training data. So, it may be also referred to as instability component. The bias component may be improved by using more complicate families of functions that are tried for data approximation. While high complexity of models often leads to increase of variance component. Such contradiction between two components is known as bias/variance dilemma [6].

Bias component structure. Let's consider structure of CCP generalized errors components. Calculation of Δ_{bias} mostly repeats calculation of $\delta(Z_{ccp})$ structure. It is evident that $\widehat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}] = \sum_{i=1}^{L} c_i \widehat{z}_i[\mathbf{x}(\omega)]$ and it is easy to show that $\{E_{\Omega}(Y \mid \mathbf{x}) - \widehat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}]\}^2$ may be written as difference

$$\sum_{i=1}^{L} c_i \{E_{\Omega}(Y \mid \mathbf{x}) - \widehat{z}_i[\mathbf{x}(\omega)]\}^2 - \sum_{i=1}^{L} c_i \{\widehat{Z}_{ccp}(\omega, \mathbf{c}) - \widehat{z}_i[\mathbf{x}(\omega)]\}^2 .$$

So,

$$\Delta_{bias}(Z_{ccp}) = \sum_{i=1}^{L} c_i E_\Omega \{E_\Omega(Y \mid \mathbf{x}) - \hat{z}_i[\mathbf{x}(\omega)]\}^2 - \sum_{i=1}^{L} c_i E_\Omega \{\widehat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}] - \hat{z}_i[\mathbf{x}(\omega)]\}^2 =$$

$$= \sum_{i=1}^{L} c_i \Delta_{bias}(z_i) - \sum_{i=1}^{L} c_i E_\Omega \{\widehat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}] - \hat{z}_i[\mathbf{x}(\omega)]\}^2 .$$

It is easy to get similar calculations from (2), showing that

$$\sum_{i=1}^{L} c_i E_\Omega \{\widehat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}] - \hat{z}_i[\mathbf{x}(\omega)]\}^2 = \sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} \rho_{i'i''}^{bc} ,$$

where $\rho_{i'i''}^{bc} = E_\Omega \{\hat{z}_{i'}[\mathbf{x}(w)] - \hat{z}_{i''}[\mathbf{x}(w)]\}^2$. Thus,

$$\Delta_{bias}(Z_{ccp}) = \sum_{i=1}^{L} c_i \Delta_{bias}(z_i) - \tfrac{1}{2} \sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} \rho_{i'i''}^{b} \qquad (4)$$

Variance component structure.

$$\Delta_{\text{var}}(Z_{ccp}) = E_\Omega E_{\Omega_t} \{\widehat{Z}_{ccp}[\mathbf{x}(\omega)] - Z_{ccp}[\mathbf{x}(\omega, \omega_t)]\}^2 =$$

$$= \sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} E_\Omega E_{\Omega_t} \{\hat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\} \{\hat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\}$$

It is evident that

$$\{\hat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\} \{\hat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\} =$$

$$= -\tfrac{1}{2} \{\hat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)] - \hat{z}_{i''}[\mathbf{x}(\omega)] + z_{i''}[\mathbf{x}(\omega, \omega_t)]\}^2 + \tfrac{1}{2} \{\hat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\}^2 +$$

$$\tfrac{1}{2} \{\hat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\}^2 .$$

However,

$$E_\Omega E_{\Omega_t} \{\hat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\}^2 = \Delta_{\text{var}}(z_{i'}) ,$$

$$E_\Omega E_{\Omega_t} \{\hat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\}^2 = \Delta_{\text{var}}(z_{i''}) .$$

Let's denote

$$z_{i'}^{vc}[\mathbf{x}(\omega, \omega_t)] = \hat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)] , \ z_{i''}^{vc}[\mathbf{x}(\omega, \omega_t)] = \hat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)] ,$$

$$\rho_{i'i''}^{vc} = E_\Omega E_{\Omega_t} \{z_{i'}^{vc}[\mathbf{x}(w, w_t)] - z_{i''}^{vc}[\mathbf{x}(w, w_t)]\}^2 .$$

Then

$$\sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} E_\Omega E_{\Omega_t} \{\hat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\} \{\hat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\} =$$

$$= \sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} \{\tfrac{1}{2}[\Delta_{\text{var}}(z_{i'}) + \Delta_{\text{var}}(z_{i''})] - \rho_{i'i''}^{vc}\} .$$

Thus,

$$\Delta_{\text{var}}(Z_{ccp}) = \sum_{i=1}^{L} c_i \Delta_{\text{var}}(z_i) - \tfrac{1}{2} \sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} \rho_{i'i''}^{vc} \qquad (5)$$

So, structure of generalized error components Δ_{bias} and Δ_{var} for CCP forecasting practically coincides with the structure of mean squared error. At that both components are always lower then corresponding components for single predictors. In other words convex combining allows to improve both contradictory constituents of bias-variance dilemma.

Variance of CCP

Let's consider structure of CCP squared variance. Let $\hat{Z}_{ccp} = E_{\Omega}(Z_{ccp})$ and $V_{ccp} = E_{\Omega}(Z_{ccp} - \hat{Z}_{ccp})^2$. Variance V_{ccp} may be written as $\sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} E_{\Omega}[\hat{z}_{i'} - z_{i'}(\omega)][\hat{z}_{i''} - z_{i''}(\omega)]$. Further calculations repeat calculations made for variance component structure evaluating:

$$[\hat{z}_{i'} - z_{i'}(\omega)][\hat{z}_{i''} - z_{i''}(\omega)] =$$

$$= -\tfrac{1}{2}[\hat{z}_{i'} - z_{i'}(\omega) - \hat{z}_{i''} + z_{i''}(\omega)]^2 + \tfrac{1}{2}[\hat{z}_{i'} - z_{i'}(\omega)]^2 + \tfrac{1}{2}[\hat{z}_{i''} - z_{i''}(\omega)]^2 .$$

Let's denote $z_{i'}^{v}(w) = \hat{z}_{i'} - z_{i''}^{v}(w)$ and $\rho_{i'i''}^{v} = E_{\Omega}[z_{i'}^{v}(w) - z_{i''}^{v}(w)]^2$. Then

$$V(Z_{ccp}) = \sum_{i=1}^{L} c_i V(z_i) - \tfrac{1}{2} \sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} \rho_{i'i''}^{v} .$$

Thus, it is shown that variance of CCP forecasting is always lower than the same convex combination of forecasting variances related to single predictors. It must be noted that decrease of prediction variance leads to loss of forecasting ability. So, additional transformation of convex forecasting function must be done with the help of uni-dimensional linear regression: $Z_{ccp}^{t} = \alpha_{ccp} Z_{ccp} + \beta_{ccp}$, where Z_{ccp}^{t} is transformed forecasting, α_t and β_t are real regression coefficients that may be found by training information with the help of least squares method. It is evident that any linear combination of predictors $\sum_{i=1}^{L} \lambda_i z_i + \gamma_0$ with $\lambda_i \geq 0$ $(i = \overline{1, L})$ may be constructed by successive execution of convex correcting and uni-dimensional linear transformation.

CCP optimization

Optimal CCP may be found by minimization of forecasting error. As it seen from expression (3) task of $\delta(Z_{ccp})$ minimization may be reduced to quadratic programming task:

$$\sum_{i=1}^{L} c_i \delta(z_i) - \tfrac{1}{2} \sum_{i'=1}^{L} \sum_{i''=1}^{L} c_{i'} c_{i''} \rho_{i'i''} -> \min \qquad (6)$$

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

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

The quadratic programming task (6) is difficult NP complete problem. But solving (6) may be facilitated with the help of procedure evaluating whether all predictors from initial set give optimal CCP or some predictors are nuisance variables and may be removed. The problem will be discussed further in terms of predictors superfluity.

Subsets $\mathbf{R}^L$, $\overline{\mathbf{D}}_L$ and $\mathbf{D}_L$ are defined as

$$\overline{\mathbf{D}}_L = \{\mathbf{c} \mid \sum_{i=1}^{L} c_i = 1, c_i \geq 0, i = \overline{1,L}\},$$

$$\mathbf{D}_L = \{\mathbf{c} \mid \sum_{i=1}^{L} c_i = 1, c_i > 0, i = \overline{1,L}\}.$$

A set of predictors will be called not superfluous or satisfying conditions of superfluity absence (CSA) if there exists a point $\mathbf{c} \in \mathbf{D}_L$ such that $\delta[Z_{ccp}(\mathbf{c})] < \delta[Z_{ccp}(\mathbf{c})]$, $\forall \mathbf{c}' \in \overline{\mathbf{D}}_L \setminus \mathbf{D}_L$.

CSA actually mean existing of CCP that uses all predictors and has forecasting error that is lower than error of any CCP that does not use all predictors. The necessary and sufficient conditions for CSA correctness that are formulated by theorem 1.

Theorem 1 . Let matrix of mutual distances between predictors $\| \rho_{i'i''} \|_{L\times L}$ is not singular and $\| \rho^{-}_{i'i''} \|_{L\times L}$ is matrix inverse to $\| \rho_{i'i''} \|_{L\times L}$. Then simultaneous correctness of inequalities

$$\sum_{i'=1}^{L} \{\delta(z_{i'})\rho^{-}_{i'i} + \frac{\frac{1}{2} - \sum_{j'=1}^{L}\sum_{j''=1}^{L} \delta(z_{j'})\rho^{-}_{j'j''}}{\sum_{j'=1}^{L}\sum_{j''=1}^{L} \rho^{-}_{j'j''}} \rho^{-}_{i'i}\} > 0$$

for $i = \overline{1,L}$ and positiveness of quadratic form $-\frac{1}{2}\sum_{j'=1}^{L}\sum_{j''=1}^{L} \rho_{j'j''}\varepsilon_{j'}\varepsilon_{j''}$ for each real vector $\varepsilon_1, \ldots, \varepsilon_L$, such that $\sum_{j=1}^{L} \varepsilon_j = 0$ is necessary and sufficient condition for CSA correctness.

Method of CCP optimization based on CSA conditions. A method for solving quadratic programming task (6) is proposed. It is based on gradual raising of predicates set meeting superfluity condition. First, a set of all possible predictor pairs P_2^{irr} is considered. A set of all irreducible pairs P_2^{irr} is then extracted using Theorem 1 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. After that an optimal aggregate is chosen, which is one from P_{i-1}^{irr} with minimum error estimate.

Experiments with CCP over uni-variate linear regressions

CCP multiple linear model. The goal of studies is performance evaluation of multiple linear regression model that is convex combination of simple regressions. At the initial stage parameters of simple linear regression models $Y = \alpha_i + \beta_i X_i + \varepsilon_i$ are evaluated by training set with least squares method (LS) for each independent variable from initial set $\tilde{X}$. So, a set of $L = |\tilde{X}|$ predictors is received:

$$\{Z_i(\omega) = \alpha_i + \beta_i X_i(\omega) \mid i = \overline{1,l}\}.$$

After that generalized errors of single predictors and discrepancies between predictors are estimated using leave-one-out technique. Then optimal CCP is searched as solution of quadratic programming task (1). Let $\mathbf{c}^0 = \{c_1, \dots, c_l\}$ are optimal CCP coefficients. That gives a solution

$$Z(\omega, \mathbf{c}^0) = \sum_{i=1}^{l} c_i \alpha_i + \sum_{i=1}^{l} c_i \beta_i X_i(\omega).$$

Usually a majority of coefficients $\mathbf{c}^0$ in high-dimensional tasks is equal to zero. So, task of CCP optimization also naturally incorporates another important task of regression analysis — significant variables selection.

CCP prognoses $Z(\omega)$ may strongly correlate with Y but at the same time forecasting errors may be great due to low variance of $Z(\omega)$. So, additional linear transformation of $Z(\omega)$ is necessary. Parameters of linear regression models $Y = \alpha_{ccp} + \beta_{ccp} X_{ccp} + \varepsilon_{ccp}$ are evaluated by training information with LS. As a result the final CCP multiple linear model is received:

$$Z(\omega, \mathbf{c}^0) = \alpha_{cpp} + \sum_{i=1}^{l} c_i \beta_{cpp} \alpha_i + \sum_{i=1}^{l} c_i \beta_{cpp} \beta_i X_i(\omega) + \varepsilon_{cpp}.$$

It must be noted that methods of regression models optimization based on quadratic programming became rather popular last years. The known Lasso technique [3] may be mentioned thereupon.

Scenarios for 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_j in j-th case is generated by formula $y_j = \sum_{k=1}^{3} u_{jk} + e_j^y$ where u_{jk} is value of latent variable U_k and e_j^y is random error term distributed $N(0,1)$.

The values of relevant variable X_i were generated by binary vector $\boldsymbol{\beta}^i = (\beta_1^i, \beta_2^i, \beta_3^i)$ In j-th case $x_{ij} = \sum_{k=1}^{3} \beta_k^i u_{jk} + e_j^{ix}$ where u_{jk} is value of latent variable U_k, $\sum_{k=1}^{3} u_{jk} = 2$, e_j^{ix} is random error term distributed $N(0, 0.05)$ The levels of irrelevant variable X_i in j-th case is generated by formula $x_{ij} = v_j^{ix}$ where v_j^{ix} is random error term distributed $N(0, d_{ix})$.

In each experiment 100 pairs of data sets were calculated by the random numbers generator according to the same scenario. The only exclusions are simulated tasks with size 50 and dimension 100. In these experiments too great amount of calculations was necessary for SR method. So only 50 pairs of data sets were generated (these results are marked asterisk in tables). Variables were selected and optimal regressions were calculated on one set from a pair and forecasting ability was evaluated on another.

First scenario. In all experiments number of relevant variables n_{rel} was fixed and equal 5: 2 were generated at $\boldsymbol{\beta} = (1,1,0)$, 2 at $\boldsymbol{\beta} = (1,0,1)$, 1 at $\boldsymbol{\beta} = (0,1,1)$. Number of irrelevant variables n_{irrel} varied and was equal 5, 20, 45, 95.

Second scenario. In experiments by this scenario number of relevant variables n_{rel} was proportional to full number of variables n_{full}. Numbers n_{full}, n_{rel} and numbers of relevant variables generated by different $\boldsymbol{\beta}$ levels are given in Table 3. Tables 4 and 5 has similar structure as Tables 1 and 2 respectively.

Table 1. Simulations parameters.

	n_{irrel}	$\boldsymbol{\beta} = (1,1,0)$	$\boldsymbol{\beta} = (1,0,1)$	$\boldsymbol{\beta} = (0,1,1)$
n_{full} = 25	13	5	5	2
nfull = 50	25	10	10	5
nfull = 100	50	20	20	10

Results. Results of experiments of the first scenario are given in Tables 1-2. In Table 1 for each pair of sample size m and full number of variables n_{full} 3 values are represented in corresponding cells: mean values of correlation coefficients between forecasted and true values of Y for CCP (upper left) and SR (upper right); fractions of tables where prognostic ability estimates for CCP regression was better than estimates for SR (bottom). In Table 2 numbers of correctly (top) and mistakenly (bottom) selected variables are represented both for CCP and SR.

Table 2. Results for the first scenario. Comparison of CCP and SR prognostic abilities.

	m = 20		m = 30		m = 50	
	CCP	SR	CCP	SR	CCP	SR
Nfull = 10	0.75	0.75	0.77	0.79	0.80	0.82
	0.43		0.30		0.36	
nfull = 25	0.78	0.64	0.78	0.72	0.79	0.77
	0.76		0.65		0.57	
nfull = 50	0.73	0.5	0.77	0.57	0.80	0.69
	0.83		0.90		0.84	
nfull = 100	0.75	0.5	0.76	0.53	0.79*	0.57*
	0.92		0.95		0.98*	

Table 3. Results of the second scenario expiriments. Numbers of correctly and mistakenly selected variables

	m = 20		m = 30		m = 50	
	CCP	SR	CCP	SR	CCP	SR
nfull = 10	235	246	258	275	290	303
	3	60	1	47	0	52
nfull = 25	236	233	255	272	287	300
	11	272	3	239	0	197
nfull = 50	227	211	259	265	279	303
	28	603	4	719	0	565
nfull = 100	218	172	244	230	139*	153*
	37	725	6	1185	0*	946*

Table 4. Results of the second scenario experiments. Comparison of CCP and SR prognostic abilities

	m = 20		m = 30		m = 50	
	CCP	SR	CCP	SR	CCP	SR
nfull = 25	0.78	0.68	0.79	0.74	0.80	0.79
	0.79		0.61		0.51	
nfull = 50	0.75	0.6	0.78	0.62	0.80	0.73
	0.82		0.87		0.78	
nfull = 100	0.75	0.5 8	0.77	0.59	0.80*	0.57*
	0.86		0.95		0.98*	

Table 5. Results of the second scenario experiments. Numbers of correctly and mistakenly selected variables.

	m = 20		m = 30		m = 50	
	CCP	SR	CCP	SR	CCP	SR
nfull = 25	253	294	288	3111	332	348
	3	171	2	156	0	120

nfull = 50	283	391	326	498	368	451
	9	335	1	397	0	307
nfull = 100	281	448	319	670	196*	529*
	11	440	2	666	0*	510*

It is seen from tables that effectiveness of SR decrease dramatically when full number of regressor variables significantly exceeds number of cases in datasets. Prognostic ability of SR decreases from 0.75-0.82 for n_{full} = 10 to 0.50-0.56 for n_{full} = 100 in first scenario experiments and from 0.78-0.79 for n_{full} = 25 to 0.57-0.59 for n_{full} = 100 in second scenario experiments. Fraction of irrelevant variables in selected set exceed 50% in all first scenario experiments with $n_{full} > 50$ At the same time CCP regression keeps efficiency in all datasets. There is only slight decrease of prognostic ability for both scenarios: from 0.75-0.80 for n_{full} = 10 to 0.75-0.795 for n_{full} = 100 in first scenario experiments and from 0.78-0.80 for n_{full} = 25 to 0.75-0.8 for n_{full} = 100 in second scenario experiments. Fraction of irrelevant variables in selected set is small in all experiments.

Conclusion

So it is shown that squared error of forecasting for CCP, CCP variance, bias and variance components of generalized error have the same structure: $\sum_{i=1}^{L} c_i t_i - \frac{1}{2}\sum_{i'=1}^{L}\sum_{i''=1}^{L} c_{i'} c_{i''} \rho^*_{i'i''}$, where t_i is corresponding term for i-th single predictor, $\rho^*_{i'i''}$ is non-negative distance function between predictors $z_{i'}$ and $z_{i''}$ that is equal 0 when predictors coincide and increase when correlation between predictors at spaces Ω or Ω_t diminishes. Thus CCP procedures allows to improve both components of bias variance decomposition. On the other hand CCP decrease also full variance of predicting functions. So additional linear transformation of CCP collective solutions is necessary.

Problem of CCP optimization was discussed. It was shown that search of optimal CCP coefficients is reduced to quadratic programming task which is solved in terms of superfluity. Concept of ensemble superfluity in CCP was discussed in details. An ensemble of predictors is called superfluous if at least one of them may be removed without loss of prediction accuracy. Necessary and sufficient conditions of superfluity absence are given in Theorem 1. A method for solving quadratic programming task using Theorem 3 has been developed. A linear regression method based on CPP optimization was considered that inherently incorporates variables selection. Testing results reveal CCP's significant superiority over stepwise regression in high-dimensional task. Method preserves effectiveness of variables selection and prognostic ability in tasks where number of potential regressor variables is several times greater than number of cases in datasets. The described results can be used in different tasks of regression analysis, forecasting or recognition.

Acknowledgment

The work was supported by Russian Foundation for Basic Research grants 08-07-00437-a, 08-01-00636-a and President's grant Scientifiic Schools 7950.2010.1.

Bibliography

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

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

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

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

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

[7] S. Geman, E. Bienenstock, R. Doursat. Neural networks and the bias/variance dilemma. NeuralComputation, 4(1):1–58, 1992.

Authors' Information

Oleg Senko *– Leading researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119333, Moscow, Vavilova, 40,* senkoov@mail.ru

Alexander Dokukin*– researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119333, Moscow, Vavilova, 40,* *dalex@ccas.ru*