The Use of Collective Method for Improvement of Regression Modeling Stability

Senko O. V.

Dorodnitsin Computer Center of Russian

Academy of Sciences

(Moscow)

email: senkoov@mail.ru

Abstract. The collective methods of regression analysis are discussed in the paper as the tools to improve the mean discrepancy of regression modeling. The mean discrepancy in the paper is calculated not only by the space of multidimensional observations but also by the space of training sets of fixed size. It was shown that mean discrepancy may be represented as the sum of three component. The first one is irremovable "noise", the second component is the mean squared deviation of mean regression function from conditional means of dependent variable in points of independent variables space, the third component is instability term that is the variance of regression functions on the space of training sets. It was shown that the using of simplest voting procedure by the initial set of regression methods allows to receive new regression model with instability component less than average instability component of the initial set .The large scale experiments with Monte-Carlo simulated data demonstrated that voting procedures really allow to improve regression performance and that the cause of such improvement is the decrease of instability of training.

Introduction

The using of collective methods for classifiers training was suggested at least 35 years ago (see Dmitriev and others (1966), Bongard (1967)) and now it is rather common in modern pattern recognition. The two main directions may be

discussed. The first one is the construction of collective solution by the set presumably trained recognition algorithms. The second direction is the search of collective solutions by the set of local regularities containing description of recognized object(see Ryazanov and others (1999), Kittler and others (1998)). In the last years the " bagging" and "boosting" methods has been developed for improving the predictive power of classifiers trained by empirical data. Both these approaches use the voting by the sets of classifiers that have been built by replicated bootstrap samples in case of "bagging" or by samples with adjusted "weights" of observations in case of "boosting" (see Brieman (1996),Shapire (1999)). On the other hand collective procedures are not so popular in regression modeling. One of the causes of such situation is heuristic character of existing collective methods and absence of statistical or mathematical grounds for their use. The approach discussed in this paper is based on minimization of mean descrepency of regression method.

2 Decomposition of the mean error of regression modeling

Let the task of prognosis of dependent variable Y by variables $X_1,\ldots,X_p$ is discussed. It is supposed that observations of variables $Y, X_1,\ldots,X_p$ are considered as elements of probability space $(\Omega_1,\Sigma_1,\mathbf{P}_1)$, where Σ_1 is Borel algebra that is generated by variables. The empirical data set $\tilde{S}$ is considered as set of independent observations from probability space $(\Omega_1,\Sigma_1,\mathbf{P}_1)$. We also may say that empirical data set $\tilde{S}$ is element of probability space $(\Omega_2,\Sigma_2,\mathbf{P}_2)$ that is direct product of m probability spaces $(\Omega_1,\Sigma_1,\mathbf{P}_1)$, where m is the number of observations in $\tilde{S}$. The optimal prognostic (regression) function are searched by empirical data set $\tilde{S}$ with the help of regression method Π. Suppose that the optimal regression function is searched in family of functions H with the help of procedure P optimizing the quality of Y approximation by $X_1,\ldots,X_p$ at data set

$\tilde{S}$. Then pair (H,P) may be considered as example of regression method. The optimal regression function that is found by data set $\tilde{S}$ with the help of regression method Π will be denoted as $h(\mathbf{x},\tilde{S},\Pi)$ or as $h(\mathbf{x},\tilde{S})$ when it is not necessary to indicate specific regression method. The data set $\tilde{S}$ that is used for the search of optimal regression function we shall call training set as it is common in pattern recognition theory.

The quality of approximation $\Delta(\tilde{S})$ of Y by regression function $h(\mathbf{x},\tilde{S})$ is naturally to estimate as mathematical mean of squared residual $\{Y-h[\mathbf{x}(\omega_1),\tilde{S}]\}^2$ or $\Delta(\tilde{S})=\int_{\Omega_1}\{Y-h[\mathbf{x}(\omega_1),\tilde{S}]\}^2\mathbf{P}_1 d(\omega_1)$

Let suppose that we want to estimate the quality of approximation of Y by regression method Π when specific training set is not known yet. In this case it is naturally to evaluate the overall quality Δ_F of regression method Π as the mean value of $\Delta(\tilde{S})$ by probability space $(\Omega_2,\Sigma_2,\mathbf{P}_2)$ or

$$\Delta_F=\int_{\Omega_2}\Delta(\omega_2)\mathbf{P}_2(d\omega_2)$$

Let $\hat{h}(\mathbf{x})$ is mathematical mean by probability space $(\Omega_2,\Sigma_2,\mathbf{P}_2)$ of regression function value at point $\mathbf{x}$: $\hat{h}(\mathbf{x})=\int_{\Omega_2}h(\mathbf{x},\omega_2)\mathbf{P}_2(d\omega_2)$.

The following theorem is true that describes the structure of Δ_F

Theorem 1.

The overall quality Δ_F of regression method Π may be represented as the sum

$$\Delta_F=\Delta_U+\Delta_I+\Delta_B, \text{ where} \qquad \textbf{(1)}$$

$$\Delta_U = \int_{\Omega_1} \{Y - \mathbf{M}[Y \mid \mathbf{x}(\omega)]\}^2 \mathbf{P}(d\omega),$$

$$\Delta_I = \int_{\Omega_2} \int_{\Omega_1} \{h[\mathbf{x}(\omega_1), \omega_2] - \hat{h}[\mathbf{x}(\omega_1)]\}^2 \mathbf{P}(d\omega_1)\mathbf{P}(d\omega_2),$$

$$\Delta_B = \int_{\Omega_1} \{\mathbf{M}[Y \mid \mathbf{x}(\omega_1)] - \hat{h}[\mathbf{x}(\omega_1)]\}^2 \mathbf{P}(d\omega_1)$$

Proof. Using simultaneous adding and extracting of terms $\hat{h}(\mathbf{x})$ and $\mathbf{M}(Y \mid \mathbf{x})$ we can represent Δ_F as

$$\Delta_F = \int_{\Omega_2} \int_{\Omega_1} \{Y(\omega_1) + \mathbf{M}[Y \mid \mathbf{x}(\omega_1)] + \hat{h}[\mathbf{x}(\omega_1)] -$$

$$-\hat{h}[\mathbf{x}(\omega_1)] - \mathbf{M}[Y \mid \mathbf{x}(\omega_1)] - h[\mathbf{x}(\omega_1), \omega_2]\}^2 \mathbf{P}_1(d\omega_1)\mathbf{P}_2(d\omega_2) \qquad (2)$$

It is seen from (2) that for proof of equality (1) it is sufficient to prove the correctness of equalities (3) and (4).

$$\int_{\Omega_1} \{Y(\omega_1) - \mathbf{M}[Y \mid \mathbf{x}(\omega_1)]\} \{\mathbf{M}[Y \mid \mathbf{x}(\omega_1)] - \hat{h}[\mathbf{x}(\omega_1)]\} \mathbf{P}(d\omega_1) = 0 \qquad (3)$$

$$\int_{\Omega_2} \int_{\Omega_1} \{Y(\omega_1) - \hat{h}[\mathbf{x}(\omega_1)]\} \times$$

$$\times \{\hat{h}[\mathbf{x}(\omega_1)] - h[\mathbf{x}(\omega_1), \omega_2]\} \mathbf{P}(d\omega_1)\mathbf{P}(d\omega_2) = 0 \qquad (4)$$

$$\int_{\Omega_2} \int_{\Omega_1} \{Y(\omega_1) - \hat{h}[\mathbf{x}(\omega_1)]\} \times$$

$$\times \{\hat{h}[\mathbf{x}(\omega_1)] - h[\mathbf{x}(\omega_1), \omega_2]\} \mathbf{P}(d\omega_1)\mathbf{P}(d\omega_2) = 0$$

The function $\hat{h}[\mathbf{x}(\omega_1)]$ is determinate function of independent variable. So it is measured function relatively independent variables vector $\mathbf{x}$. The function $\theta[\mathbf{x}(\omega_1)] = \mathbf{M}[Y \mid \mathbf{x}(\omega_1)] - \hat{h}[\mathbf{x}(\omega_1)]$ is also measured function as it is sum of two measured functions. Using the properties of measured functions we can receive that

$$\mathbf{M}[Y\theta(\mathbf{x}) \mid \mathbf{x}] = \theta(\mathbf{x})\mathbf{M}(Y \mid \mathbf{x}) \tag{5}$$

By definition of conditional mathematical mean

$$\int_{\Omega_1} \{Y(\omega_1)\theta[\mathbf{x}(\omega_1)] - \mathbf{M}[Y\theta \mid \mathbf{x}(\omega_1)]\}\mathbf{P}(d\omega_1) = 0 \tag{6}$$

The equality (3) follows directly from equalities (5) and (6).

The equality (4)) follows from definition of $\hat{h}(\mathbf{x})$ and Fubini theorem. ||

Let discuss components Δ_U, Δ_I, Δ_B and the ways of their reducing . The conditional mean $\mathbf{M}(Y \mid \mathbf{x})$ is the best prognosis of Y at point $\mathbf{x}$. So the term Δ_U is the minimal achieved deviation of Y from prognoses by the given set of prognostic variables. So the quality of prognosis may be improved only by minimization of the terms Δ_I and Δ_B. The term Δ_B describe the quality of prognoses of regression function $\hat{h}(\mathbf{x})$ that is mean regression function by the space of all possible training sets. The main approach for Δ_B reducing is using of the family H that contains regression functions more close to $\mathbf{M}(Y \mid \mathbf{x})$. The instability term Δ_I depends on the complexity of family H, the size and structure of data set $\widetilde{S}$. The most simple way to diminish Δ_I is of course expanding of training sets at fixed complexity level of family H. But in many practical task such possibility is only hypothetical. The alternative way is using instead of family H the family of less complicated regression functions. Today there are many approaches exist to find the complexity level that is adequate to existing training set size. The most popular method is use of Akaike Informational

Criterion and related criteria (see). The fundamental approach to the problem was suggested by Vapnik (1982).

The factor that may affect Δ_I is existence of outlying observation randomly scattered in data. Outliers are inconsistent with the main bulk of data and so they may randomly deviate the regression surface . In this case Δ_I can be decreased by using robust optimization procedures that are based on downweighting of outlying observation or full ignoring them as the most radical approach.

3 The collective methods as the tool to improve the regression stability and exactness.

This paper is focused at approach for Δ_I decreasing that is based on collective procedures. Let suppose that the set of regression method is $\tilde{\Pi} = \{\Pi_1, \ldots, \Pi_L\}$ used in some prognostic task. Let $\tilde{h} = \{h_1[\mathbf{x}, \tilde{S}], \ldots, h_L[\mathbf{x}, \tilde{S}]\}$ is the set of corresponding optimal regression functions. In case when there is no sufficient reasons to presume one of the methods the collective solution can be used. The most simple collective method is the use the mean by the set $\tilde{h}$ regression function $h_a[\mathbf{x}, \tilde{S}, \tilde{h}] = \frac{1}{L}\sum_{i=1}^{L} h_i[\mathbf{x}, \tilde{S}]$. The collective method generating regression function $h_a[\mathbf{x}, \tilde{S}, \tilde{h}]$ will be referred to as the Method of Normal Values (MNV) and denoted as $\Pi_a(\tilde{\Pi})$. Let consider the instability component of method $\Pi_a(\tilde{\Pi})$. Let $\hat{h}_a[\mathbf{x}, \tilde{h}] = \int_{\Omega_2} h_a[\mathbf{x}, \omega_2, \tilde{h}]\mathbf{P}_2(d\omega_2)$, then

$$\Delta_I[\Pi_a(\tilde{\Pi})] = \int_{\Omega_2}\int_{\Omega_1} \{h_a[\mathbf{x}(\omega_1), \omega_2, \tilde{h}] - \hat{h}_a[\mathbf{x}(\omega_1), \tilde{h}]\}^2 \mathbf{P}_1(d\omega_1)\mathbf{P}_2(d\omega_2)$$

The variability $V(\Pi,\mathbf{x})$ at point $\mathbf{x}$ of the method Π generating optimal regression function $h(\mathbf{x})$ is defined as integrand $\int_{\Omega_2}\{h[\mathbf{x},\omega_2]-\hat{h}[\mathbf{x}]\}^2\mathbf{P}_2(d\omega_2)$.

The following Lemma will be necessary in following discussions.

Lemma. For arbitrary $p\in\{1,\ldots,\infty\}$ and arbitrary set of real numbers $\{a_1,\ldots,a_p\}$ the inequality $p[\sum_{j=1}^{p}a^2{}_j]\geq[\sum_{j=1}^{p}a_j]^2$ is correct .

Proof. It is sufficient to show that quadric form $\Phi(a_1,\ldots,a_p)=p[\sum_{j=1}^{p}a^2{}_j]-[\sum_{j=1}^{p}a_j]^2$ is nonnegative or it is sufficient to show that eigenvalues of corresponding to Φ matrix $\mathrm{M}_\Phi=\begin{pmatrix} p-1 & -1 & . & . & -1\\ -1 & p-1 & . & . & -1\\ . & . & . & . & .\\ -1 & . & . & p-1 & -1\\ -1 & . & . & -1 & p-1\end{pmatrix}_{p\times p}$ are nonnegative. The matrix a M_Φ may be represented as the sum $\mathrm{M}_\Phi=p\mathbf{I}_{p\times p}+\mathrm{M}^{-1}_{p\times p}$, where $\mathbf{I}_{p\times p}$-identity matrix and $\mathrm{M}^{-1}_{p\times p}$ is the matrix with all elements equal -1. Matrix $\mathrm{M}^{-1}_{p\times p}$ has two eigenvalues 0 and $-p$. So the matrix M_Φ has two eigenvalues 0 and p.||

Theorem 2. For instability component of collective method $\Pi_a(\tilde{\Pi})$ the unequality $\Delta_I[\Pi_a(\tilde{\Pi})]\leq\frac{1}{L}\sum_{i=1}^{L}\Delta_I(\Pi_i)$ is true.

Proof. The ratio

$$\kappa_{ij}(\mathbf{x})=\frac{\int_{\Omega_2}\{h_i[\mathbf{x},\omega_2]-\hat{h}_i[\mathbf{x}]\}\times\{h_j[\mathbf{x},\omega_2]-\hat{h}_j(\mathbf{x})\}\mathbf{P}_2(d\omega_2)}{\sqrt{\int_{\Omega_2}\{h_i[\mathbf{x},\omega_2]-\hat{h}_i(\mathbf{x})\}^2\mathbf{P}_2(d\omega_2)}\sqrt{\int_{\Omega_2}\{h_j[\mathbf{x},\omega_2]-\hat{h}_j(\mathbf{x})\}^2\mathbf{P}_2(d\omega_2)}}$$

belongs to the cut [0,1], because it is correlation coefficient of two random functions.

The instability component can be represented as

$$\Delta_I[\Pi_a(\tilde{\Pi})] = \int_{\Omega_1} \{\frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\int_{\Omega_2} \{h_i[\mathbf{x}(\omega_1),\omega_2] - \hat{h}_i[\mathbf{x}(\omega_1)]\}\times$$

$$\times\{h_j[\mathbf{x}(\omega_1),\omega_2] - \hat{h}_j[\mathbf{x}(\omega_1)]\}\mathbf{P}_2(d\omega_2)\}\mathbf{P}_1(d\omega_1) =$$

$$= \int_{\Omega_1} \{\frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\kappa_{ij}(\mathbf{x})\sqrt{V(\Pi_i,\mathbf{x})}\ \sqrt{V(\Pi_j,\mathbf{x})}\}\mathbf{P}_1(d\omega_1) \le$$

$$\le \int_{\Omega_1} \{\frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\sqrt{V(\Pi_i,\mathbf{x})}\ \sqrt{V(\Pi_j,\mathbf{x})}\}\mathbf{P}_1(d\omega_1) =$$

$$= \int_{\Omega_1} \frac{1}{L^2}[\sum_{j=1}^{L}\sqrt{V(\Pi_i,\mathbf{x})}]^2\mathbf{P}_1(d\omega_1)$$

Further using Lemma we receive

$$\int_{\Omega_1} \frac{1}{L^2}[\sum_{j=1}^{L}\sqrt{V(\Pi_i,\mathbf{x})}]^2\mathbf{P}_1(d\omega_1) \le \int_{\Omega_1} \frac{1}{L}[\sum_{j=1}^{L}V(\Pi_i,\mathbf{x})]\mathbf{P}_1(d\omega_1)$$

Taking into account that by definition of variability $\int_{\Omega_1} V(\Pi_i,\mathbf{x})\mathbf{P}_1(d\omega_1) = \Delta_I(\Pi_i)$

we receive inequality $\Delta_I[\Pi_a(\tilde{\Pi})] \le \frac{1}{L}\sum_{i=1}^{L}\Delta_I(\Pi_i).\|$

As it follows from the proved theorem 2 the use of even the most simple of collective methods (MNV) lead to better stability than the mean stability of the initial set of regression methods. So the use of the MNV method may lead to better prognoses than prognoses of methods from initial set. The improvement is of course possible if the shift component of collective solution does not increase too much. The existence of such possibility is demonstrated by simulation results. The method MNV implying the equal include of regression methods from initial set is not the single way to construct final collective rule.

The natural generalization of MNV is collective method calculating by set regression $\tilde{h} = \{h_1[\mathbf{x}, \tilde{S}], \ldots, h_L[\mathbf{x}, \tilde{S}]\}$ regression function $h_w(\mathbf{x}, \tilde{S}, \tilde{\Pi}, \mathbf{c}) = \sum_{i=1}^{L} c_i h_i(\mathbf{x}, \tilde{S})$ where $c_1, \ldots, c_L$ are nonnegative weighing parameters satisfying condition $\sum_{i=1}^{L} c_i = 1$. One of the way to find the optimal values of $c_1, \ldots, c_L$ is instability term minimization. . The instability component may be represented as

$$\int_{\Omega_1} \{\frac{1}{L^2} \sum_{i=1}^{L} \sum_{j=1}^{L} c_i c_j \kappa_{ij}(\mathbf{x}) \sqrt{V(\Pi_i, \mathbf{x})} \sqrt{V(\Pi_j, \mathbf{x})}\} \mathbf{P}_1(d\omega_1) = \frac{1}{L^2} \sum_{i=1}^{L} \sum_{j=1}^{L} c_i c_j \gamma_{ij},$$

where $\gamma_{ij} = \int_{\Omega_1} \kappa_{ij}(\mathbf{x}) \sqrt{V(\Pi_i, \mathbf{x})} \sqrt{V(\Pi_j, \mathbf{x})} \mathbf{P}_1(d\omega_1)$. So the task of instability component minimization may be reduced to the bilinear programming task:

$$\frac{1}{L^2} \sum_{i=1}^{L} \sum_{j=1}^{L} c_i c_j \gamma_{ij} \to \min$$

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

The main weakness of the weighing parameters evaluating by instability component minimization is the full ignoring of the shift component behavior. The

another drawback is the necessity of sufficiently exact estimates of coefficients γ_{ij}. The represented below experiments with simulated data sets demonstrate that sufficiently simple heuristic methods of the regression methods from initial set $\tilde{\Pi}$ weighing may be successful.

The method of weighed pair regressions (WPR)

In method of weighed paired regressions the initial set of regression methods $\tilde{\Pi}_2$ includes all methods using as specific prognostic variables two independent variables from the set $X_1,\ldots,X_p$. In methods from $\tilde{\Pi}_2$ optimal regression functions are linear functions that are found by the least squares method. Let $\{h_{ij}^p(x_i,x_j,\tilde{S})\,|\,i,j\in I_p\}$ is the set of optimal regression functions corresponding to regression methods from $\tilde{\Pi}_2$ calculated by training set $\tilde{S}=\{(y_1,\mathbf{x}_1),\ \ldots\ ,(y_m,\mathbf{x}_m)\}$, $I_p=\{1,\ldots,p\}$. The $\hat{d}_{ij}(\tilde{S})$ is square discrepancy of the optimal regression function $h_{ij}^p(x_i,x_j,\tilde{S})$ calculated by training set $\tilde{S}$ itself or $\hat{d}_{ij}(\tilde{S})=\frac{1}{m}\sum_{j=1}^{m}[y_j-\hat{\boldsymbol{\beta}}\mathbf{x}_j^{+1}]^2$ where $\mathbf{x}_j^{+1}$ is expansion of vector of regressor variables $\mathbf{x}_j$ from training set $\tilde{S}$ and $\hat{\boldsymbol{\beta}}$ is the vector of regression parameters calculated by $\tilde{S}$. Let $\hat{d}_{min}(\tilde{S})$ is the minimal square discrepancy by all methods from $\tilde{\Pi}_2$. The parameter of **WPR** method is threshold for model selection that is selected from interval (0,1). The collective regression function $h_{wp}(\mathbf{x},\kappa,\tilde{S})$ in WPR method is calculated with the help of weighed voting procedure by all optimal regression functions from $\{h_{ij}^p(x_i,x_j,\tilde{S})\,|\,i,j\in I_p\}$ for which inequality $\frac{1}{\hat{d}_{ij}(\tilde{S})}>\kappa\frac{1}{\hat{d}_{min}(\tilde{S})}$ is correct. Let $\tilde{H}_s$ is the set of such functions then

$$h_{wp}(\mathbf{x},\kappa,\tilde{S}) = \frac{\sum\limits_{h^p{}_{ij}\in\tilde{R}_S} \frac{1}{d_{ij}(\tilde{S})} h_{ij}^p(x_i,x_j,\tilde{S})}{\sum\limits_{h_{ij}^p\in\tilde{R}_S} \frac{1}{d_{ij}(\tilde{S})}}$$

4. Simulations

The Monte-Carlo simulation was used to evaluate the performance of voting procedures in regression tasks. The $m_{ts} = 1000$ pairs of equal size data sets were generated in each study. In each pair one data set was used as training set and another(control set) was used to evaluate the exactness of prognoses. The number of observations in training (control) set will be denoted as m. The last one will be called control set. Besides in each study the 300 vectors of X variables were generated by the same procedure as they were generated in training and control sets. This set of vectors will be referred as $\widetilde{X}_{ins}$ and its elements will be denotes as $\mathbf{x}_j^s = (x_{1j}^s,\ldots,x_{pj}^s)$. This set of vectors was used to estimate the instability term in examined regression methods.

Let $\hat{\boldsymbol{\beta}}(\tilde{s}_i^t)$ - the vector of regression parameters calculated by training set $\tilde{s}_i^t$. The value of discrepancy δ_i^r of the model received by training set $\tilde{s}_i^t$ was calculated by control set $\tilde{s}_i^c = \{(y_{i1}^c,\mathbf{x}_{i1}^c),\ldots,(y_{im}^c,\mathbf{x}_{im}^c)\}$: $\delta_i^r = \sum\limits_{j=1}^{m}[y_{ij}^c - \hat{\boldsymbol{\beta}}_i(\tilde{s}_i^t)\mathbf{x}_{ij}^{c+1}]^2$, where $\mathbf{x}_{ij}^{c+1} = (1,x_{ij1}^c,\ldots,x_{ijp}^c)$. The overall relative discrepancy (quality of modeling) is evaluated by ratio $\Delta_F^e = \dfrac{\sum\limits_{i=1}^{\mathrm{m}_t}\delta_i^r}{\sum\limits_{i=1}^{\mathrm{m}_t}\sum\limits_{j=1}^{m}(y_{ij}^c-\bar{y})^2}$, where $\bar{y} = \frac{1}{\mathrm{m}_t m}\sum\limits_{i=1}^{\mathrm{m}_t}\sum\limits_{j=1}^{m} y_{ij}^c$. The value

of overall relative instability of modeling is evaluated by $\widetilde{X}_{ins}$ as

$$\Delta_I^e = \frac{\sum_{j=1}^{m_{ins}} \sum_{i=1}^{m_t} [\hat{\boldsymbol{\beta}}(\tilde{s}_i^t) - \bar{\boldsymbol{\beta}}]\mathbf{x}_j^{c+1}}{\sum_{i=1}^{m_t} \sum_{j=1}^{m} (y_{ij}^c - \bar{y})^2}, \text{ where } \bar{\boldsymbol{\beta}} = \frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\boldsymbol{\beta}}(\tilde{s}_i^t) \text{ and } \mathbf{x}_j^{s+1} = (1, x_{j1}^s, \ \dots \ , x_{jp}^s)$$

The two scenarios of simulation experiments were used that correspond two considered types of dependence of Y on X variables. In each study the performance of methods based voting procedures was compared with standard multiple linear regression with coefficients found by least square method.

The first scenario. The vector levels of X variables are independent distributed multivariate normal with mean 7.5 and standard deviation 4.0. The Y value in the *i*-th case is generated by $y_i = x_i\boldsymbol{\beta} + e_i$ where x_i is the vector of levels of X variables, $\boldsymbol{\beta}$ is vector of regression coefficients and e_i is random error term distributed $N(0,d)$. The values of regression coefficients belong to $\{0,5\}$.

The second scenario. The regressor variables X so as dependent variable Y are the stochastic functions of two latent variables Z_1 and Z_2. The vector levels of Z variables are independent distributed multivariate normal with mean 0 and standard deviation 1.0. The value x_{ji} of variable X_j in *i*-th case is generated by $x_{ij} = \sum_{k=1}^{2} z_{ik}\gamma_{kj} + e(d_{xj})$ where z_{ik} is corresponding value of latent variable γ_{kj} - real parameter, $e(d_{xj})$ is random error term distributed $N(0,d_{xj})$. The value of dependent variable y_i in *i*-th case is generated by $y_i = \sum_{k=1}^{2} z_{ik}\chi_k + e(d_y)$ where z_{ik} the value of latent variable Z_k, χ_k is real parameter, is random error term distributed $N(0,d_y)$.

The evaluating MNV performance. The initial set of regression methods includes methods that were based on different groups of independent variables. The data sets were generated according second scenario. The 16 independent variables were partitioned on four equal groups. Inside each group the 2 variables were generated with γ_{kj} parameters values equal 0 and another 2 variables were generated with γ_{kj} parameters values equal 5.0. In other words the 2 variables in each group were "noisy" variables not related with response. The dependent variable was generated with χ_k parameters values equal 5.0. The results of experiments are represented in table 1.

Table 1

	Group 1 Standard Regression.	Group 2 Standard Regression.	Group 3 Standard Regression.	Group 4 Standard Regression.	MNV (16 var.)	Standard Regression. (16 var.)
Δ_F^e	0.257	0.259	0.259	0.259	0.192	0.270
Δ_I^e	0.037	0.038	0.04	0.04	0.0144	0.09

It is seen that stability of MNV is significantly better than stability of standard regression for the task with 16 variables so as for all tasks with 4 variables. The re improvement of exactness also exists but it is not so great of course. It must be noted that for task with 16 variables improvement of exactness in absolute terms is rather close to improvement of stability.

The evaluating of the WPR method performance. The WPR method was compared with standard regression in experiments that were generated according first (Table 2) and second scenarios (Table 3). The number of observations in data sets m, the full number p of independent variables X, the number p_l of independent variables that are not related with response, the variance of error term

of response d. In each study the mean relative discrepancy Δ_F^e and the mean relative instability Δ_I^e.

Table 2 First scenario.

	m	p	p_l	d	Δ_F^e	. Δ_I^e
WPR	40	8	0	12	0.695	0.199
St.Reg.	40	8	0	12	0.575	0.128
WPR	40	16	0	12	0.675	0.251
St. Reg.	40	16	0	12	0.476	0.20
WPR	40	16	8	12	0.747	0.231
St. Reg.	40	16	8	12	0.730	0.307
WPR	40	20	10	12	0.785	0.266
St.Reg.	40	20	10	12	0.806	0.420
WPR	70	16	8	12	0.611	0.148
St. Reg.	70	16	8	12	0.549	0.131
WPR	40	32	16	12	0.836	0.312
St. Reg.	40	32	16	12	1.81	1.53
WPR	60	32	16	12	0.654	0.252
St. Reg.	60	32	16	12	0.625	0.348
WPR	50	32	16	12	0.747	0.267
St. Reg.	50	32	16	12	0.855	0.579
WPR	70	32	16	12	0.644	0.261
St. Reg.	70	32	16	12	0.523	0.249
WPR	50	32	0	12	0.741	0.288
St. Reg.	50	32	0	12	0.493	0.334

Table 3. Second scenario.

	n	p	p_l		Δ_F^e	. Δ_I^e
WPR	70	16	8	12	0.761	0.024
St.Reg.	70	16	8	12	0.975	0.234
ВПР	70	16	8	6	0.453	0.015
St. Reg.	70	16	8	6	0.578	0.139
WPR	70	8	4	6	0.413	0.014
St.Reg.	70	8	4	6	0.459	0.055
WPR	70	8	0	6	0.4	0.012
St. Reg.	70	8	0	6	0.445	0.053
WPR	70	8	0	12	0.684	0.02
St. Reg.	70	8	0	12	0.763	0.09
WPR	50	8	0	12	0.7	0.028
St.Reg.	50	8	0	12	0.82	0.134
WPR	40	8	0	12	0.712	0.037
St.Reg.	40	8	0	12	0.886	0.186
WPR	40	8	4	12	0.727	0.044
St.Reg	40	8	4	12	0.897	0.189
WPR	30	8	4	12	0.746	0.065
St.Reg.	30	8	4	12	0.996	0.285
WPR	70	18	0	18	0.736	0.02
St.Reg.	70	18	0	18	0.996	0.273

Discussion of simulations results. The results of experiments with data sets that has been generated according the first scenario demonstrate significantly better performance of standard regression method for all studies where the number of observations in data sets are more than two times greater than the number of independent variables. For majority of such experiments the stability of standard multiple regression is also better than stability of WPR. The results can be easily explained because the model that was used for data sets generating completely coincides with the linear function that is searched in standard multiple regression . However the exactness of WPR method is sometimes better standard multiple regression in studies where data sets have been generated according the first scenario but the number of observation is closer to number of regressor variables. It is seen from the table that better exactness of WPR is achieved due to better stability.

As distinct from the first scenario the all represented in table 2 results of studies with data sets generated according the second scenario demonstrate the better performance of WPR method. It is seen from the table also that the difference between exactness of standard multiple regression and WPR is close to the difference between instability components of two methods. So we may conclude that the cause of better performance of MNV in these studies is its better stability.

5. Conclusions

The experiments with Monte-Carlo simulated data confirmed the theoretical result that was formulated in theorem 2 that collective solution by the set of recognition methods allows to receive the new technique with better stability than the average stability of methods from initial set. The experiments also demonstrated that the increase of stability leads to improving of regression exactness. The large scale experiments with heuristic MPR technique demonstrated the good performance of

this method in tasks where the serious problems with stability exist i.e. in high-dimensional task with limited number of observations.

The research was supported by Russian Fond of Basic Researches 03-01-00580 and INTAS 00-626.

References

Dmitriev A.N., Yu.I. Zhuravlev , Krendelev F.P. About Mathematical Principles of Objects and Phenomena Classification "Discreet Analysis", v.7, Nauka, Siberian Department, Novosibirsk, 1966 (in Russian)..

Bongard M.M. The problem of recognition. Moscow, Nauka, 1967(in Russian).

Chatfield, C. Model Selection, Data Mining and model Uncertainty.18th International Workshop on Statistical Modelling. July 7-11, 2003, Leuven, Belgium, p.79-84.

Vapnik V.N.(1982). *Estimation of Dependencies Based on Empirical data.* New York: Springer.

V.V.Ryazanov, O.V.Senko, and Yu.I. Zhuravlev. Methods of recognition and prediction based on voting procedures. //Pattern Recognition and Image Analysis, Vol. 9,N 4, 1999, p.713-718.

J.Kittler, M.Hatef, R.P.W.Duin, J. Matas. On combining classifiers.//IEEE Transactions on Pattern Analysis and Machine Intelligence.-1998,N 20, p.226-239.

Brieman L.(1996) Bagging predictors. *Machine Learnig* **24**, 123-140.

R.E. Shapire. A Brief Introduction to Boosting. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. 1999.