Forecasting using predictor selection from a large set of highly correlated variables

A Yu Timofeeva[1] and Yu A Mezentsev[1]

[1]Novosibirsk State Technical University, K. Marksa Avenue, 20,Novosibirsk, Russia 630073

e-mail: a.timofeeva@corp.nstu.ru

Abstract. The potential of correlation-based feature selection has been explored in selecting an optimal subset from a set of highly correlated predictors. This problem occurs, for example, in time series forecasting of economic indicators using regression models on multiple lags of a large number of candidate leading indicators. Greedy algorithms (forward selection and backward elimination) in such cases fail. To obtain the globally optimal solution, the feature selection problem is formulated as a mixed integer programming problem. To solve it, we use the binary cut-and-branch method. The results of simulation studies demonstrate the advantage of using the binary cut-and-branch method in comparison with heuristic search algorithms. The real example of the selection of leading indicators of consumer price index growth shows the acceptability of using the correlation-based feature selection method.

1. Introduction

Big data analytics includes the feature selection task [1, 2] for predictive modelling [3]. In many practical applications, candidate predictors correlate strongly. An example is the task of time series forecasting using leading indicators [4].

Lagged predictors are highly correlated. Fast and scalable univariate feature selection methods are not suitable under a given situation. They evaluate features individually, so the final subset includes many redundant strongly correlated features.

Multivariate methods take into account feature dependencies and try to discard not only irrelevant variables (which do not affect the response), but also redundant ones. Most often, the predictors are selected simultaneously with the construction of predictive models using embedded methods such as LASSO regression [4, 5]. It provides a sparse solution that includes only relevant features, which, however, is very sensitive to the regularization parameter.

In addition, stepwise regression is often used in time series forecasting [6]. It refers to the so-called "wrapper" methods. They select the optimal subset of features from all possible candidates simultaneously with the model estimation. Generally, this problem has exponential complexity in the number of features. In practice, to solve it, search approaches use greedy algorithms [7]. However, they do not usually produce an optimal solution, but approximate a globally optimal solution in a reasonable amount of time.

Finally, filter methods select variables regardless of the model. A Correlation-based Feature Selection is a well-known multivariate filter algorithm [8]. This approach is proposed for solving

classification problems. Its applicability for predictor selection with a highly correlation of candidate predictors, in particular when selecting leading indicators, is poorly studied. This is the gap that we will attempt to address in our article. For this purpose, we first transform a correlation-based heuristic evaluation function optimization problem into a mixed integer programming problem. But in this problem the number of variables and constraints depends on the square of the number of features. With a branch and bound algorithm, the amount of computation becomes large. Therefore it is proposed to use the previously developed binary cut-and-branch method.

2. Correlation-based Feature Selection

Correlation-based Feature Selection (CFS) ranks feature subsets according to a correlation-based heuristic evaluation function [8]. The best subset contains predictors highly correlated with the response, yet uncorrelated to each other. Thus, the problem of feature selection is formulated as the following optimization problem:

$$\frac{\sum_{i \in S_k} R_i}{\sqrt{k + 2 \sum_{i,j \in S_k, i \neq j} r_{ij}}} \to \max_{S_k}, \tag{1}$$

where R_i is an absolute value of correlation coefficient between the response and the *i*-th feature, r_{ij} is an absolute value of correlation coefficient between the *i*-th and the *j*-th features, S_k is a subset of *k* features.

As for the time series of economic indicators, both the response and the predictors are usually quantitative. Therefore the Pearson product-moment correlation coefficient is applicable.

We reformulate the problem (1) as a problem of nonlinear integer programming:

$$\frac{\sum_{i-1}^{n} R_i^2 x_i + 2 \sum_{i \neq j} R_i R_j x_i x_j}{\sum_{i-1}^{n} x_i + 2 \sum_{i \neq j} r_{ij} x_i x_j} \to \max_{x_1, \ldots, x_n}, \tag{2}$$

where $x_i \in \{0,1\}, i = 1,...,n$, n is the number of features. If $x_i = 1$ then the optimal subset contains the *i*-th variable, and $x_i = 0$, otherwise.

The problem (2) is a polynomial fractional programming problem. Based on the transformation of feature selection problem proposed in [9], we replace the denominator in (2) by a positive continuous variable *u*. It leads to the equivalent polynomial problem. In addition, we convert a maximization problem into a minimization one. Thus, the problem is represented as follows:

$$-\sum_{i-1}^{n} R_i^2 x_i u - 2 \sum_{i \neq j} R_i R_j x_i x_j u \to \min_{x_1, \ldots, x_n, u}$$

$$\sum_{i-1}^{n} x_i u + 2 \sum_{i \neq j} r_{ij} x_i x_j u = 1,$$

$$u > 0,\ x_i \in \{0,1\}.$$

Based on a linearization technique proposed in [10] to transform the terms $x_i u$, $x_i x_j u$, we introduce variables z_i, $i = 1,...,n$, v_{ij}, $i = 1,...,n$, $j = 1,...,n$, $i \neq j$. Then we obtain the following mixed integer linear programming problem:

$$-\sum_{i-1}^{n} R_i^2 z_i - 2 \sum_{i \neq j} R_i R_j v_{ij} \to \min_{\substack{x_1, \ldots, x_n, u, \\ z_1, \ldots, z_n, v_{11}, \ldots, v_{nn}}} \tag{3}$$

$$z_i \geq 0,\ v_{ij} \geq 0,\ u > 0,\ x_i \in \{0,1\},$$

$$\sum_{i-1}^{n} z_i + 2 \sum_{i \neq j} r_{ij} v_{ij} = 1,$$

$$M\left(x_i-1\right)+u\le z_i\le M\left(1-x_i\right)+u,\ z_i\le Mx_i,$$
$$M\left(x_i+x_j-2\right)+u\le v_{ij}\le M\left(2-x_i-x_j\right)+u,\ v_{ij}\le Mx_i, v_{ij}\le Mx_j,$$

where M is a large positive value.

Finally, the initial nonlinear problem (1) is reduced to a high-dimensional linear programming (LP) problem. The number of the new continuous variables is $(n^2+n+2)/2$. They are added to the initial n binary variables. The number of constraints also depends on n^2 and is $(2n^2+n+2)$.

3. Binary cut-and-branch method

The binary cut and branch method (BCBM) was originally developed to solve LP problems with Boolean variables [11] and then extended to the case of the General linear programming problem with mixed variables (milp) [12]. Any such problem, the special case of which is (3), can be represented as:

$$\gamma(x)=c^{1T}x+c^{2T}y+const\to\max, \tag{4}$$
$$A^1x+A^2y\le b,\ \bar{0}\le x\le\bar{1},\ y\ge\bar{\bar{0}}, \tag{5}$$
$$x\in I_2^{n^1}, \tag{6}$$

which is a milp problem with Boolean variables x and continuous variables y. Conditions (3) specify that the solution components x belong to one of the vertices of the unit hypercube of dimension n^1; $c^1, x, \bar{0}, \bar{1}$ are vectors of the same dimension; $\bar{0}$ is a zero vector; $\bar{1}$ is a vector of ones; *const* is a constant; and $c^1\ge\bar{0}$. The vectors $c^2, y, \bar{\bar{0}}$ have dimension n^2. Condition $x\in I_k^n$ indicates that x belongs to the set of vectors of dimension n, the elements of which take integer values from the range $[0\div k-1]$. Conformity of statements (4-6) and (3): Boolean variables $x\in I_2^{n^1}$ have the same meaning in both statements, continuous variables $y\ge\bar{\bar{0}}$ have the meaning of variables $z_i\ge 0, v_{ij}\ge 0, u>0$ in statement (3). Matrices A^1 and A^2 in the constraints (4) are formed from the coefficients of the constraints in the problem (3).

In fact, any milp problem and a considerable part of mip problems can be compactly reduced to (1)–(3); see, e.g., [11].

Suppose x^0, y^0 is the solution of the relaxed problem (4)–(5); $[\cdot]$ is the integer part of number; and $\beta_0=\hat{\alpha}^Tx^0$, where $\hat{\alpha}_j\in\{0,\ 1\}, j=\overline{1,n}$. Then any inequality of the form

$$\hat{\alpha}^Tx\le\hat{\beta}_0,\ \hat{\alpha}_j\in\{0,\ 1\},\ j=\overline{1,n},\ \hat{\beta}_0=[\beta_0],\ \beta_0=\hat{\alpha}^Tx^0, \tag{7}$$

is called a binary cut (BC) for problem (4)–(5).

If x^0 is part of the solution x^0, y^0 of the relaxed problem (4)–(5), then

$$\varsigma^Tx\le\phi_0,\ \phi_0=\varsigma^Tx^0. \tag{8}$$

Relation (8) can be a generating inequality if $\varsigma_j=\sum_{i\in I^B}\lambda_ia_{ij}$, $\lambda_i\ge 0$, where $a_{ij}, i\in I^B$ are the coefficients of the basis part A^1 and λ_i are the weights of the basis constraints. Specifically, if λ_i are dual estimates for constraints (5), then $\varsigma_j=c_j$, $j=\overline{1,n}$. A complementary system of BCs to constraints (5) is defined as

$$\mathrm{A}^{1D}x\le\beta, \tag{9}$$

where $\mathrm{A}^{1D}=\left\|\hat{\alpha}_j^i\right\|_{m^D\times n^1}$, $\hat{\alpha}_j^i\in\{0,\ 1\}, j=\overline{1,n}$ is the coefficient matrix of the complementary system and the vector β composed of the right-hand part of the constraints is defined from (7).

There are several ways to find out whether the BCs are valid [9,10]. Specifically, the binary cut and branch algorithm (BCBA) uses the following feature. We now define

$$\hat{\alpha}^T x \geq \beta(x^0)+1, \tag{10}$$

where $\beta(x^0)=\left[\hat{\alpha}^T x^0\right]$ and x^0 is part of the optimal solution x^0, y^0 of problem (4), (5), and (9).

If problem (4), (5), and (10) has a solution, then the cut $\hat{\alpha}^T x \leq \beta(x^0)$ is invalid. Contrary wise, if problem (4), (5), and (10) has no solution due to the conflicting conditions (5) and (10), then $\hat{\alpha}^T x \leq \beta(x^0)$ is a valid BC. This is the underlying feature of the BC synthesis procedure called *selection in a set of the nearest cuts* (SSNC) [11, 12]. We now describe this procedure.

We define an inequality ensuing from the basis system (5) and (9) through a permutation by arranging ς in a descending order (denoted by $\vec{\varsigma}$). We consider a totality of n^1 vectors, of dimension n^1:

$$\hat{\alpha}^1=(1,0,...,0),\ ...,\ \hat{\alpha}^j=(1,1,...,1,0,0,...,0)\ (\ j\ \text{original ones}),\ ...,\ \hat{\alpha}^{n^1}=(1,1,...,1).$$

Each $\hat{\alpha}^j$ is set in correspondence with the value $cs(\hat{\alpha}^j)=\dfrac{\vec{\varsigma}^T\hat{\alpha}^j}{|\varsigma|_2\left|\hat{\alpha}^j\right|_2}$, $j=\overline{1,n^1}$.

The discrete function $cs(\hat{\alpha}^j)$ has a strict maximum and uniquely defines the priority of each of the alternative cuts with the coefficients $\hat{\alpha}^j$. Adding the entire totality of these BCs to (9) and solving (4), (5), and (10), we can find out conflicting conditions (if there are any) to identify valid cuts. Then, if there are valid BCs, we select a single cut with the maximum value of $cs(\hat{\alpha}^j)$. If there are no valid cuts, we select a BC that corresponds to the maximum of $cs(\hat{\alpha}^j)$, $j=\overline{1,n^1}$.

Another important feature of BCs is their radicality measure, which characterizes the depth of a cut of a given type. For a BC $\hat{\alpha}^T x\leq\hat{\beta}_0$, $\hat{\alpha}_j\in\{0,1\}$, $j=\overline{1,n}$, $\hat{\beta}_0=[\beta_0]$, $\beta_0=\hat{\alpha}^T x^0$, we define the radicality r as the number of vertices of the unit hypercube cut off by the BC (BC system), assuming that the cut is valid.

In the general case, $\hat{a}^T x\leq b$, $b\in I_k^1, x\in I_2^{n^1}, \hat{a}\in I_2^{n^1}$, $k=\sum_{j=1}^{n^1}\hat{a}_j, 1\leq k\leq n^1$ or $\sum_{j=1}^{n^1}\hat{a}_j x_j\leq b, x_j\in I_2^1, j=\overline{1,n^1}$, where $\hat{a}_j\in\{0,1\}$ are the cut-off coefficients.

For an arbitrary $b\in I_1^k$, $k=\sum_{j=1}^{n^1}\hat{a}_j, 1\leq k\leq n^1$, and $C_k^l=\dfrac{k!}{l!(k-l)!}$, we define

$$r_k^b=2^{n^1-k}\sum_{l=b+1}^{k}C_k^l\to\max \tag{11}$$

Relation (11) considers the unit hypercube vertices lying above the level $\hat{a}^T x\leq b$, i.e., belonging to the hyperplanes $\hat{a}^T x=l$ with the right-hand parts ($b+1, b+2, ..., k$). The $\hat{a}^T x=b$ hyperplane itself contains $2^{n^1-k}C_k^b$ vertices. The *maximum radical BC* is derived from (11) and $\hat{\alpha}_j=1, j=\overline{1,n^1}$ with the possible exclusion of the minimum order relative to $\vec{\varsigma}$ if the sum of the coefficients in the left-hand part of the BC $\hat{a}^T\tilde{x}\leq b$ without this exclusion is an integer number.

Regardless of which measure—closeness to the generating inequality or radicality is considered as a priority, the BCBA is as follows.

3.1. Binary Cut-and-Branch Algorithm

1. Suppose that we have obtained the solution of the original relaxed problem (4), (5), and (7): x^0, y^0 and $\gamma(x^0, y^0)$. If x^0 are integers, the algorithm stops. Otherwise, it goes to step 2.

2. At step $t(1,2,...)$, we select a probing vertex with the maximum estimate $\gamma(x^q, y^q)$, $q \in (1,2,...,t-1)$. If the list of vertices is empty, the problem has no integer solution. The algorithm stops. If the vertex with the maximum estimate $\gamma(x^q, y^q)$ contains integer x^q, the solution (x^q, y^q) is the optimal one. The algorithm stops. Otherwise:

3. We create two new candidates for each of which we supplement the current matrix A^{1D} for the step q with BCs (7) and (10) by the cut selection procedures (by the value of $cs(\hat{\alpha}^j)$ or by radicality (11)): $(\hat{\alpha}^{(t+1)})^T x \le \beta(x^q)$ и $(\hat{\alpha}^{(t+1)})^T x \ge \beta(x^q)+1$, respectively.

4. We solve a pair of alternative subproblems with the cuts $(\hat{\alpha}^{(t+1)})^T x \le \beta(x^q)$ and $(\hat{\alpha}^{(t+1)})^T x \ge \beta(x^q)+1$.

5. We save their solution components $\underline{x}^{t+1}$ and $\bar{x}^{t+1}$ and the estimates $\gamma(\underline{x}^{t+1}, \underline{y}^{t+1})$ and $\gamma(\bar{x}^{t+1}, \bar{y}^{t+1})$ by adding them to a list of the tree vertices. If any of the candidates has no solution, it is withdrawn from the list of the vertices.

6. We increase the step number ($t := t+1$) and go to step 2.

4. Simulation study results

The applicability of the CFS method to variable selection with a highly correlation of candidate predictors was investigated using the following model example.

The relevant features $x_1^{(m)}, x_2^{(m)}, x_3^{(m)}, x_5^{(m)}, x_6^{(m)}$ were modeled as independent random variables with a standard normal distribution. The relevant features $x_4^{(m)}, x_7^{(m)}$ were computed as follows:

$$x_4^{(m)} = x_3^{(m)} + e_1, x_7^{(m)} = x_5^{(m)} + x_6^{(m)} + e_2,$$

where e_1, e_2 are independently standard normally distributed.

The response was defined as

$$y = \sum_{i=1}^{7} x_i^{(m)} + e_3$$

where e_3 is a normally distributed random variable independent of e_1, e_2 and of $x_i^{(m)}$ with zero expectation and standard deviation equal to 0.1.

The redundant variables were modeled correlated with the main predictors:

$$x_i^{(r)} = x_i^{(m)} + \varepsilon_i,\ i = 1,...,7,$$
$$x_i^{(r)} = x_{i-7}^{(m)} + \varepsilon_i,\ i = 8,...,14,$$

where $\varepsilon_1,...,\varepsilon_{14}$ are independently standard normally distributed. The irrelevant features $\xi_1,...,\xi_5$ were also modeled as random noise. In addition, the noise candidates include $\varepsilon_1,...,\varepsilon_4$.

For each random variable, samples of size 1,000 were drawn. As a result, the set of features for selection included

- the relevant predictors, on the basis of which the response was calculated, $x_1^{(m)},...,x_7^{(m)}$;
- the redundant variables correlated with the relevant one, $x_1^{(r)},...,x_{14}^{(r)}$;
- the irrelevant features $\xi_1,...,\xi_5$, $\varepsilon_1,...,\varepsilon_4$.

A total number of variables was thirty. Simulation studies were repeated 1,000 times. The mixed integer linear programming problem contained 496 variables and 1,832 constraints. In order to solve the problem, greedy algorithms were used as an alternative: forward selection and backward elimination [13]. Their implementation in the R environment was used: the forward.search and

backward.search functions provided by the FSelector package. The variable subset was evaluated by the objective function from (1).

Table 1 presents the results of simulation studies, the proportion of cases in which each variable is included in the subset. The irrelevant features were not included in the subset in any experiment. The following notation is used: forward — the forward selection, backward — the backward elimination, BCBM — the binary cut-and-branch method.

Table 1. Predictor selection results.

	Method				Method		
Variable	forward	backward	BCBM	Variable	forward	backward	BCBM
$x_1^{(m)}$	0.219	0.955	0.927	$x_4^{(r)}$	0.049	0.938	0.613
$x_2^{(m)}$	0.216	0.945	0.929	$x_5^{(r)}$	0.006	0.716	0.024
$x_3^{(m)}$	0.258	0.991	0.962	$x_6^{(r)}$	0.013	0.742	0.036
$x_4^{(m)}$	0.954	0.996	0.98	$x_7^{(r)}$	0.217	0.997	0.915
$x_5^{(m)}$	0.208	0.993	0.977	$x_8^{(r)}$	0	0.046	0.006
$x_6^{(m)}$	0.204	0.995	0.969	$x_9^{(r)}$	0	0.061	0.003
$x_7^{(m)}$	1	1	0.999	$x_{10}^{(r)}$	0	0.069	0.005
$x_1^{(r)}$	0.011	0.462	0.057	$x_{11}^{(r)}$	0.012	0.37	0.053
$x_2^{(r)}$	0.013	0.504	0.066	$x_{12}^{(r)}$	0.001	0.065	0.002
$x_3^{(r)}$	0.012	0.642	0.053	$x_{13}^{(r)}$	0	0.048	0.001

From table 1 it can be seen that the forward selection often does not include the redundant predictors. However, the final subset almost does not get a significant part of the relevant features. Only $x_4^{(m)}, x_7^{(m)}$ correlated with the other significant predictors are constantly included in the subset. This has a negative effect on the objective function values, which are far from optimal. The evaluation function values are displayed as a boxplot in figure 1. From figure 1, it can be concluded that the variation of the objective function values in the simulations is very large, and the average value is much smaller than that achieved by the backward elimination and the binary cut-and-branch method.

Both the backward elimination and the binary cut-and-branch method include all relevant features in the subset of predictors. But the backward elimination often leaves too many redundant predictors in the subset. Compared to this, for the BCBM results, such cases are relatively rare. Only a few redundant attributes $x_4^{(r)}, x_7^{(r)}$ are often included in the subset. This is because they are related to the relevant predictors $x_4^{(m)}, x_7^{(m)}$ which are correlated with the other significant predictors. Nevertheless, despite this problem, the binary binary cut-and-branch method provides the best values of the objective function compared to the backward elimination (figure 1).

The multicollinearity problem described is evidently difficult for the CFS method. It is a problem of the method itself, rather than optimization algorithms. This may have a negative effect on its applicability in practice.

5. Selection of leading indicators of consumer price index growth

Let us verify the applicability of the Correlation-based Feature Selection method using a real example. For this purpose, the task of forecasting the consumer price index was chosen. From economic research [14] it is known that one of the leading indicators of price changes during a business cycle is the industrial materials price index. The data provided by the unified interdepartmental informational-statistical system [15] include the monthly time series of the price index for the acquisition of machines and equipment for investment purposes. Indices are grouped by type of economic activity and regions of the Russian Federation. Further, the territory of the Russian Federation as a whole is selected.

The period from June to November 2010 was chosen as the test time interval. During that period, there was a sharp increase in the base price index in percent compared to the corresponding period of the previous year. The training dataset was taken from January 2006 to May 2010. It was used to fit the ARIMA model. But it naturally predicts the continuing fall in prices, as shown in figure 2. Let us verify whether this forecast can be improved by using leading indicators.

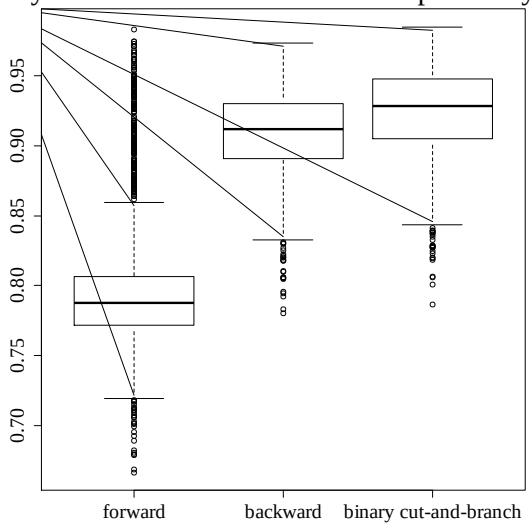

Figure 1. Boxplot for the values of the objective function obtained in simulation studies.

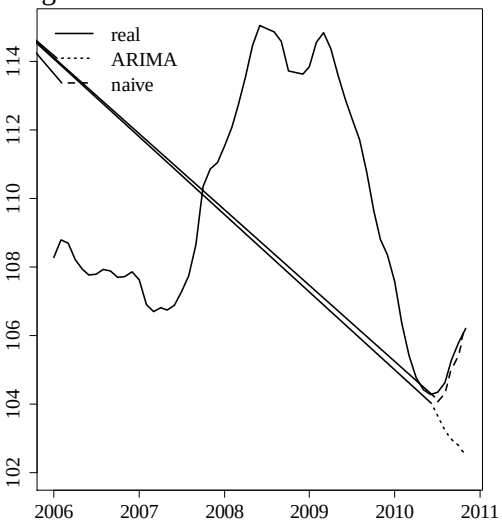

Figure 2. Real and forecast values of the base consumer price index, in percent compared to the corresponding period of the previous year.

For the time period used to train the model, there are the data on price indices for the acquisition of machines and equipment for investment purposes for 92 types of economic activity, including 16 sections of Classification and a total index for all types of activity. The values of indices are in percent compared to the corresponding period of the previous year.

First, a naive approach was used, suggesting that the best predictor is the total index for all activities. The CFS method was used for optimal lags selection only. Lags from 0 to -6 were considered as candidates. As a result, lags -4 and -6 are chosen. This approach led to a rather low value of the objective function F^* (table 2).

Next, the price indices for the acquisition of machines and equipment for investment purposes by economic activity, taken with lags from 0 to -6, were considered as candidate predictors for the application of greedy algorithms. Thus, the total number of predictors n was 644. The obtained optimal number n^* is presented in table 2.

Table 2. The results of solving the optimization problem.

Method	n	n^*	F^*
naive	7	2	0.857957
forward selection	644	4	0.954952
backward elimination	644	39	0.963451
forward selection	119	4	0.942901
backward elimination	119	18	0.950590
binary cut-and-branch	119	7	0.954994

Finally, the number of predictors was reduced to indices by sections and the total index, that is, 17 indices. The total number of variables was 119 taking into account the possible lags. Table 2 shows the optimal values of the objective functions F^* and the number of predictors n^* obtained using the methods of forward selection, backward elimination, and binary cut-and-branch.

In both cases, the forward selection method leaves very few predictors. The values of the objective function are the lowest. The method of backward elimination selects a lot of variables, which, of course, are redundant. The binary cut-and-branch method provides the best value of the objective function.

In order to compare the forecasting performance on test data, regression models were estimated on the training set with the inclusion of selected variables with selected lags. For this, the dynlm package of the statistical environment R was used.

The future predictor values should be available in order to build a forecast for six months ahead. It was assumed that their real values are available only until May 2010. If later values were needed, then the predictor values were forecasted based on the ARIMA model. For the automatic selection of the structure of the ARIMA model, the auto.arima function of the forecast package was used. It is implemented in the R environment. For forecasting, the forecast function from the same package was used. Forecasting based on the results of the predictor selection by the backward elimination method was not performed, since the number of variables (39 and 18) is clearly redundant for estimating a model of 47 months (taking into account the earliest lag -6).

Table 3 presents the deviations of the real values of the consumer price index from its forecasts. The smallest absolute differences are in bold. It is revealed that the naive approach gives a good result. Graphically, it is shown in figure 2.

Table 3. Consumer price index forecast results.

Method	June	July	August	September	October	November
ARIMA	0.22	0.68	1.39	2.31	2.98	3.73
naive	**-0,07**	0.28	0.30	0.27	**0.36**	**-0.18**
forward selection, 92 indices	-1,6	-0.9	-0.67	**0.18**	0.71	0.6
forward selection, 17 indices	-0.08	0.53	0.37	0.84	1.06	0.69
binary cut-and-branch, 17 indices	-0.86	**0.13**	**0.29**	0.93	1.22	1.13

This means that the Correlation-based Feature Selection approach can be recommended for choosing the optimal lags of predictor variables in a time series model. The simultaneous selection of price indices for the acquisition of machines and equipment for investment purposes by economic activity and their lags is more complicated. The time series of indices are very similar (some are even almost identical). Hence there is a very high correlation between the variables. At the same time, the lagged values of the indices are also highly correlated. This is similar to the case of multicollinearity from the model example considered in the previous section. As it was revealed above, in such a model, the achievement of optimum in problem (1) does not guarantee that only relevant features will be selected. In the structure of the solution, a certain proportion of redundant predictors is allowed.

Evidently, this problem also occurs in the selection of indices as leading indicators. As a result, the optimal solution obtained using the binary cut-and-branch method provides the best forecast performance for the medium term (2-3 months) only. In the long term, 5-6 months ahead, forecasts are worse than when selecting optimal lags for the total index only.

This effect is clearly visible when compared the results of selection by forward selection and binary cut-and-branch method for 17 indices. The results of forward selection give a smaller number of predictors and better predict for the long term. This can be explained by the overfitting effect, since using the binary cut-and-branch method, a greater number of predictors have been selected, some of which may be redundant.

6. Conclusions and recommendations

Thus, the Correlation-based Feature Selection method is applicable for selecting leading indicators in the time series forecasting. The candidate predictors are highly correlated. From simulation studies, it is revealed that greedy heuristics optimization algorithms in such cases do not give satisfactory results.

The forward selection does not include many relevant features; the backward elimination leaves many redundant predictors in the subset. The binary cut-and-branch method gives the best result. However, the CFS method is not perfect: the optimal value of the heuristics does not guarantee that only all relevant predictors will be selected, and the inclusion of redundant variables is possible. This happens when the relevant features correlate with each other and with redundant variables. To avoid this, it is recommended by forming the initial set of candidate predictors to pre-exclude duplicate indicators and indicators with very similar dynamics.

7. References

[1] Gaidel A V and Krasheninnikov V R 2016 Feature selection for diagnosing the osteoporosis by femoral neck X-ray images *Computer Optics* **40(6)** 939-946 DOI: 10.18287/2412-6179-2016-40-6-939-946
[2] Kutikova V V and Gaidel A V 2015 Study of informative feature selection approaches for the texture image recognition problem using the Laws' masks *Computer Optics* **39(5)** 744-750 DOI: 10.18287/0134-2452-2015-39-5-744-750
[3] Bolón-Canedo V, Sánchez-Maroño N and Alonso-Betanzos A 2015 *Knowledge-Based Systems* **86** 33-45
[4] Sagaert Y R, Aghezzaf E H, Kourentzes N and Desmet B 2018 *European Journal of Operational Research* **264** 558-569
[5] Tibshirani R 1996 *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267-288
[6] Fite J T, Don Taylor G, Usher J S, English J R and Roberts J N 2002 *International Journal of Physical Distribution & Logistics Management* **32** 299-308
[7] Flach P 2012 *Machine learning: the art and science of algorithms that make sense of data* (Cambridge University Press)
[8] Hall M A 1999 *Correlation-based feature selection for machine learning* PhD thesis (Hamilton: University of Waikato)
[9] Nguyen H, Franke K and Petrovic S 2009 *Proc. of the NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Sub modularity, Sparsity & Polyhedra (DISCML)* (Vancouver, Canada)
[10] Chang C-T 2001 *European Journal of Operational Research* **131** 224-227
[11] Mezentsev Y A 2016 Binary Cut-and-Branch Method for Solving Linear Programming Problems with Boolean Variables *CEUR Workshop Proceedings* **1623** 72-85
[12] Mezentsev Y 2017 *Constructive Nonsmooth Analysis and Related Topics (dedicated to the memory of V.F. Demyanov) (CNSA)* (St. Petersburg, Russia) 1-3
[13] Sutter J M and Kalivas J H 1993 *Microchemical journal* **47** 60-66
[14] Klein P A and Moore G H 1983 *Journal of forecasting* **2** 119-135
[15] Unified interdepartmental informational-statistical system URL: https://fedstat.ru/

Acknowledgments

The reported study was funded by Russian Ministry of Education and Science, according to the research project No. 2.2327.2017/4.6.