Raport Badawczy

Research Report

RB/47/2011

Feature selection based on the AIC helpfulness criterion

M. Ostrycharz, P. Grzegorzewski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6
01-447 Warszawa
tel.: (+48) (22) 3810100
fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. zw. dr hab. inż. Olgierd Hryniewicz

Warszawa 2011

Feature selection based on the AIC helpfulness criterion

Małgorzata Ostrycharz[1], **Przemysław Grzegorzewski**[1,2]

[1]Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warsaw, Poland
[2]Faculty of Mathematics and Computer Science,
Warsaw University of Technology
Plac Politechniki 1, 00-661 Warsaw, Poland
malgorzata.ostrycharz@gmail.com, pgrzeg@ibspan.waw.pl

Abstract

Importance of feature selection techniques in multidimensional data analysis is nowadays beyond doubt. It is especially so in such learning tasks which are characterized by a very high dimensionality and a low number of learning examples. An alternative approach to well known and commonly used selection methods (e.g. backward, forward, stepwise) is to use the Akaike Information Criterion ($\mathcal{AIC}$) for feature selection investigating the whole feature set simultaneously.

An experimental approach to feature selection suggested in the paper is based on so-called AIC Improvement Matrices, which describe the situation in the whole feature set. Besides paying attention to AIC selection algorithms refer also to correlation between features in the data set.

Keywords: AIC, Akaike criterion, feature selection, data sets, data mining.

1 Introduction

One of the key problems in data mining is to search the best approximating model g such that

$$Y_i = g(x_{i,1}, x_{i,2}, \ldots, x_{i,p}, \varepsilon_i), \tag{1}$$

where Y is the dependent variable, $x_1, x_2, \ldots, x_p$ are realizations of random explanatory variables $X_1, X_2, \ldots, X_p$, ε is a random factor and $i = 1, 2, \ldots, n,$.

We assume that function g depends on some parameter $\underline{\theta}$. We can consider many approximation functions, each parameterized by some $\underline{\theta}$ from the possible parameter space $\underline{\Theta}$. Since we have got features to describe their influence on the dependent random variable Y (the target), we can think of selecting the best approximating model g in terms of selecting **the best feature set** to describe the target [4]. Thus each parameter $\underline{\theta}$ denotes some candidate feature set. However feature selection is not only an optimal choice of one feature set in order to describe the target in the best possible way. We should be aware that each modelling often faces problems of complexity, executability and significance.

In this paper we propose an **experimental** approach to select a few most important features which influence the target. Our method is a competitive to popular forward, backward and stepwise selection methods. It is based on the **Akaike information criterion** $\mathcal{AIC}$ ([1]). Taking into account accuracy and complexity models with the lowest $\mathcal{AIC}$ indicator are supposed to have good predictive properties (see [5]).

Although formulated on the ground of information theory, the Akaike information criterion is applied for different tasks, like state-space model selection [2], problems related to time series and regression [6], ensemble neural networks [10], etc. Some improvements of the original AIC criterion were also proposed, e.g. its bootstrap variant [9].

We try to cope with both - relevance to the target and predictive accuracy - to deal with large data sets as well as controlling relations between features.

The paper is organized as follows: In Sec. 2 we describe the main idea of the contribution. Next, in Sec. 3, we describe the so-called $\mathcal{AIC}$ matrices and explain their possible usefulness in variable selection. Then we present our main algorithm (Sec. 4) which is later illustrated on the leukemia data set (Sec. 5).

2 The objective

Let us assume a data set with n observations with a target Y (being continuous or discrete, e.g. binary) and q continuous variables features) given by a matrix X_q. We can build many models with the dependent variable Y and descriptive variables from the X_q. Under established criteria one can select an optimal model and hence obtain some features defining this model. According to common methodology of feature selection we can use the model accuracy as the performance measure (so called wrapper method). Therefore we select the model with the highest predictive accuracy and regard the features used by this model as the **optimal features**. Unfortunately each wrapper introduces its own bias when estimating the accuracy, i.e. why a wrapper taken to features selection determines the type of the model to be finally trained [8]. Moreover, in a large data set processing wrappers may be quite difficult or even impossible due to the wrapper's handicap of handling high dimensional data.

The objective is to look at **interactions** between features and simultaneously pay attention to descriptive properties of more than one feature to the target. The idea is to select the best features according to *predictive accuracy improvement* which results from adding a second feature to a single feature model.

Here we use the $\mathcal{AIC}$ measure as a measure of a model's predictive accuracy. Therefore we will use regression to provide $\mathcal{AIC}$ indicator. The approach implies that the number of features q in the model should be less than the number of observations n.

3 $\mathcal{AIC}$ Matrices

Let us adopt the following notation: m_{ij} stands for a regression model based on the $y \sim v_i + v_j$ formula, m_i denotes a regression model based on the $y \sim v_i$ formula and m_0 stands for a regression model where a model is described only by an intercept (i.e. $y \sim 1$ formula).

Let us consider a matrix of simple regression models each with a single descriptive feature or intercept only (Tab. 1). We call this matrix *basic model matrix*.

	v_1	v_2	$\dots$	v_{n-1}	v_n
v_1	m_0	m_1	$\dots$	m_1	m_1
v_2	m_2	m_0	$\dots$	m_2	m_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
v_{n-1}	m_{n-1}	m_{n-1}	$\dots$	m_0	m_{n-1}
v_n	m_n	m_n	$\dots$	m_n	m_0

Table 1: Basic model matrix

In this case a variable connected with a **row** is called the **basic variable** and a model with this variable and the intercept is built in each cell corresponding to this row, apart from the cells on the diagonal which are filled by simple *target* $\sim$ *intercept* models. Now we add a variable corresponding to a **column** to each model existing in the *basic model matrix.* We receive the so-called *full model matrix* (Tab. 2).

	v_1	v_2	...	v_{n-1}	v_n
v_1	m_1	$m_{1,2}$	...	$m_{1,n-1}$	$m_{1,n}$
v_2	$m_{2,1}$	m_2	...	$m_{2,n-1}$	$m_{2,n}$
...	...	...	...	...	...
v_{n-1}	$m_{n-1,1}$	$m_{n-1,2}$	...	m_{n-1}	$m_{n-1,n}$
v_n	$m_{n,1}$	$m_{n,2}$	...	$m_{n,n-1}$	m_n

Table 2: Full model matrix

The *full model matrix* is symmetrical since model $m_{i,j}$ is identical with $m_{j,i}$ for each i, j, $i \neq j$.

Let us calculate the $\mathcal{AIC}$ over the proposed models. Assume that $\mathcal{AIC}_{ij}$ stands for an $\mathcal{AIC}$ of a regression model based on the $y \sim v_i + v_j$ formula; $\mathcal{AIC}_i$ denotes $\mathcal{AIC}$ of a regression model based on the $y \sim v_i$ formula and $\mathcal{AIC}_0$ equals to $\mathcal{AIC}$ for the regression model based on the $y \sim 1$ formula.

The $\mathcal{AIC}$ matrix corresponding to *full model matrix* will be called a *full $\mathcal{AIC}$ matrix*, whereas $\mathcal{AIC}$ matrix corresponding to *basic model matrix* will be called a *basic $\mathcal{AIC}$ matrix.* Both matrices are presented in Table 3 and 4.

	v_1	v_2	...	v_{n-1}	v_n
v_1	$\mathcal{AIC}_1$	$\mathcal{AIC}_{1,2}$	...	$\mathcal{AIC}_{1,n-1}$	$\mathcal{AIC}_{1,n}$
v_2	$\mathcal{AIC}_{2,1}$	$\mathcal{AIC}_2$	...	$\mathcal{AIC}_{2,n-1}$	$\mathcal{AIC}_{2,n}$
...	...	...	...	...	...
v_{n-1}	$\mathcal{AIC}_{n-1,1}$	$\mathcal{AIC}_{n-1,2}$	...	$\mathcal{AIC}_{n-1}$	$\mathcal{AIC}_{n-1,n}$
v_n	$\mathcal{AIC}_{n,1}$	$\mathcal{AIC}_{n,2}$	...	$\mathcal{AIC}_{n,n-1}$	$\mathcal{AIC}_n$

Table 3: $\mathcal{AIC}$ for full model matrix

Each $\mathcal{AIC}$ index tells us how "good" is the model. In particular, $\mathcal{AIC}_i$ gives information about a "predictive goodness" of feature v_i, $\mathcal{AIC}_0$ tells about a predictive accuracy of the model with the intercept, which is a

	v_1	v_2	$\ldots$	v_{n-1}	v_n
v_1	$\mathcal{AIC}_0$	$\mathcal{AIC}_1$	$\ldots$	$\mathcal{AIC}_1$	$\mathcal{AIC}_1$
v_2	$\mathcal{AIC}_2$	$\mathcal{AIC}_0$	$\ldots$	$\mathcal{AIC}_2$	$\mathcal{AIC}_2$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
v_{n-1}	$\mathcal{AIC}_{n-1}$	$\mathcal{AIC}_{n-1}$	$\ldots$	$\mathcal{AIC}_0$	$\mathcal{AIC}_{n-1}$
v_n	$\mathcal{AIC}_n$	$\mathcal{AIC}_n$	$\ldots$	$\mathcal{AIC}_n$	$\mathcal{AIC}_0$

Table 4: $\mathcal{AIC}$ for basic model matrix

mean of a target, whereas $\mathcal{AIC}_{i,j}$ shows how good are predictive properties of features v_i and v_j in describing the dependent variable.

If after adding a variable to the existing model $\mathcal{AIC}$ of the new model gets lower, then it means that this new variable has a positive impact on describing the target together with the existing variable. We can say this new variable "helps" the previous variable. The lower is the new $\mathcal{AIC}$ (in comparison to the $\mathcal{AIC}$ of the previous model), the better *predictive progress* has been made. Taking the foregoing into account we can consider two kinds of $\mathcal{AIC}$ **predictive accuracy improvement**:

- absolute improvement

$$\mathcal{AI}^{\mathcal{AIC}} = \mathcal{AIC}(basic\ model) - \mathcal{AIC}(full\ model), \tag{2}$$

- relative improvement

$$\mathcal{RI}^{\mathcal{AIC}} = \frac{\mathcal{AIC}(basic\ model) - \mathcal{AIC}(full\ model)}{\mathcal{AIC}(basic\ model)}. \tag{3}$$

Let us consider a matrix $\mathcal{AI}^{\mathcal{AIC}}$ containing the following elements:

$$\mathcal{AI}^{\mathcal{AIC}}_{i,j} = \mathcal{AIC}_i - \mathcal{AIC}_{i,j} \quad \text{for } i \neq j$$

and

$$\mathcal{AI}^{\mathcal{AIC}}_{i,i} = \mathcal{AIC}_i - \mathcal{AIC}_0.$$

Analogously, we can construct a matrix $\mathcal{RI}^{\mathcal{AIC}}$. If given nondiagonal element in $\mathcal{AI}^{\mathcal{AIC}}$ or $\mathcal{RI}^{\mathcal{AIC}}$ is positive then variables corresponding to that element's coordinates describe the target better than a single variable corresponding to the row coordinate. According to the sign of diagonal elements we can distinguish two kinds of variables:

- if $\mathcal{AI}^{AIC}_{i,i}$ (or $\mathcal{RI}^{AIC}_{i,i}$) is **positive** then we will call v_i a **strong variable,**
- if $\mathcal{AI}^{AIC}_{i,i}$ (or $\mathcal{RI}^{AIC}_{i,i}$) is **negative** then we will call v_i a **weak variable.**

A strong variable can describe the target better than only an intercept, whereas a weak variable cannot do this. However, it is possible that a weak variable describes the target fairly well together with other variable.

Experimental methods are based on a conjecture that if each two features from some feature set F have good predictive accuracy, then F might have good predictive accuracy too. For example, having $F = \{v_i, v_j, v_k\}$ with $\mathcal{RI}^{AIC}$ being positive for each pair, we may expect that a model based on the features from F would have good predictive accuracy.

4 Filtering the most helpful features

Since the data size is huge (thousands or hundreds of features) a model cannot be directly applied to the whole data set. We pay attention on the $\mathcal{AIC}$ improvement as a general result of developing model with two variables in comparison with the model with a sinle variable. We came up with that methodology not only for the reason of a very good and intuitive matrix representation of two variable models, but the conjecture that it is possible to "approximate" the predictive accuracy of multivariable model by predictive accuracy of many bivariable models. Being more precise, if we consider many bivariable models on the basis of some feature set (every combination of two features appears) which all have a positive predictive accuracy improvement (adding second feature to the first lower the $\mathcal{AIC}$), the multivariable model based on the whole feature set may present good predictive accuracy too.

The algorithm goes forward starting from a single feature and adding another features. Thus an important question arises: Which feature to choose as the starting one? It is so important because the first feature selection determines (at least to a certain degree) further variables. Hence the first feature should be chosen as good as possible both with respect to the target and to other features.

Below we propose three selection methods:

1. the best column sum in $\mathcal{AI}^{AIC}$ (or $\mathcal{RI}^{AIC}$) matrix;

2. the best column sum in $\mathcal{AI}^{\mathcal{AIC}}$ (or $\mathcal{RI}^{\mathcal{AIC}}$) matrix among only those columns which have the maximal number of positive elements;

3. the best element on the diagonal.

Using the first approach we indicate a feature which generates the highest usefulness in all feature sets. The second method starts from the preselection of features that can bring profit to the highest number of all other features (positive elements), and then marks out the most desired one (the best column sum). According to the third approach we simply choose the strongest feature among all available.

Let us adopt the following notation: FF will denotes a *final feature set* (i.e. a set containing finally selected features), FL will stand for a *features left set* (i.e. a set of available features we have at the beginning of each step of the algorithm) and FC will stand for a *features candidate set* (i.e. a set containing candidates to FF).

Now we are able to present the main algorithm for selecting features based on $\mathcal{AIC}$ improvement:

Algorithm

1. Select the first feature (using any method described above) and add this feature to FF. Mark the initial feature set without the first feature by FL.

2. Repeat until FL is empty:

   - Choose feature candidates into FC as those features f_c in FL which fulfill the following condition

   $$\left(\mathcal{RI}^{\mathcal{AIC}}_{f,f_c} > 0, \mathcal{RI}^{\mathcal{AIC}}_{f_c,f} > 0\right) \quad \forall (f \in FF). \tag{4}$$

   - Compute the weight of each feature candidate $f_c \in FC$ either as

   $$W^{\mathcal{RI}}_{f_c,FF} = \sum_{f \in FF} \left(\mathcal{RI}^{\mathcal{AIC}}_{f,f_c} + \mathcal{RI}^{\mathcal{AIC}}_{f_c,f}\right), \tag{5}$$

   or as

   $$W^{\mathcal{RI}}_{f_c,FF} = \sum_{f \in FF} \mathcal{RI}^{\mathcal{AIC}}_{f,f_c}. \tag{6}$$

- Take such $\tilde{f}_c \in FC$ which has the highest candidate weight $W^{\mathcal{RI}}_{\tilde{f}_c,FF}$ and add it to FF.
- Update FL set by excluding the selected feature $\tilde{f}_c$ from FC, i.e.
$$FL := FC \setminus \left\{\tilde{f}_c\right\}.$$

3. Return FF as the final feature set.

Please note, that we may consider a similar algorithm taking $\mathcal{AI}^{\mathcal{AIC}}$ instead of $\mathcal{RI}^{\mathcal{AIC}}$.

5 Illustrative example

For better understanding let us consider the following example.

Example

Consider a logistic regression model for a leukemia data set with the following features:

$$leukemia.exp = \{g48,\ g49,\ g50,\ g65,\ g88,\ g92,\ g98,\ g112,\ g133,$$
$$g134,\ g136,\ g139\}$$

and a binary target Y.

$\mathcal{RI}^{\mathcal{AIC}}$	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}
v_1	0,08	-0,012	0,003	0,248	0,201	0,088	0,052	0,067	0,037	0,180	0,040	0,175
v_2	-0,007	0,080	-0,002	0,237	0,250	0,070	0,089	0,061	0,054	0,157	0,053	0,233
v_3	-0,012	-0,023	0,098	0,233	0,229	0,058	0,077	0,055	0,036	0,168	0,057	0,238
v_4	0,139	0,123	0,136	0,200	0,257	-0,026	0,124	0,036	0,000	0,022	0,029	0,173
v_5	0,035	0,090	0,083	0,216	0,242	0,209	0,029	0,059	0,046	0,148	0,005	0,042
v_6	0,081	0,058	0,065	0,096	0,340	0,092	0,169	0,072	0,029	0,078	0,071	0,167
v_7	0,029	0,063	0,070	0,216	0,177	0,156	0,106	0,070	0,037	0,152	0,006	0,101
v_8	0,072	0,060	0,074	0,161	0,224	0,084	0,096	0,080	0,079	0,122	0,040	0,122
v_9	0,055	0,067	0,068	0,142	0,225	0,054	0,077	0,092	0,067	0,096	0,052	0,189
v_{10}	0,124	0,095	0,124	0,086	0,246	0,022	0,115	0,058	0,015	0,143	0,058	0,199
v_{11}	0,051	0,060	0,082	0,162	0,187	0,089	0,041	0,048	0,046	0,129	0,073	0,098
v_{12}	0,130	0,187	0,209	0,237	0,164	0,128	0,074	0,070	0,129	0,209	0,037	0,132

Table 5: leukemia.exp.AIC.perc.matrix

Using a mapping

$$g48 \to v_1,\ g49 \to v_2, \dots,\ g136 \to v_{11},\ g139 \to v_{12},$$

we get a matrix $\mathcal{RI}^{\mathcal{AIC}}$ (Tab. 5). Now we can calculate some useful measures, like:

- Sum of $\mathcal{RI}^{\mathcal{AIC}}$ in columns (AIC.wgt.sum.col):

$\mathcal{RI}^{\mathcal{AIC}}$	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}
	0.781	0.848	1.011	2.233	2.742	1.024	1.049	0.767	0.577	1.604	0.521	1.867

- Sum of $\mathcal{RI}^{\mathcal{AIC}}$ in rows (AIC.wgt.sum.row):

$\mathcal{RI}^{\mathcal{AIC}}$	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}
	1.163	1.274	1.216	1.213	1.203	1.318	1.182	1.214	1.184	1.286	1.066	1.704

We can interpret AIC.wgt.sum.col[i] as showing "how much variable v_i is helpful to other variables" while AIC.wgt.sum.row[i] tells us "how much other variables help variable v_i". We will select the first feature according to the best "AIC.wgt.sum.col" and add the feature to the *final feature set* FF, i.e.

$$FF := \{v_5\}\,.$$

We obtain the *feature left set* FL as the initial feature set without this variable. Since we have chosen the first feature, we start the main algorithm.

1^{st} selection

- We choose feature candidates into FC as those features f_c from FL, which fulfill (4):

$$FC := \{v_1, v_2, v_3, v_4, v_6, v_7, v_8, v_9, v_{10}, v_{11}, v_{12}\}$$

- We compute the weight for each feature candidate $f_c \in FC$ using formula (6)

$W^{\mathcal{RI}}_{f_c,(v_5)}$	v_1	v_2	v_3	v_4	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}
	0.035	0.090	0.083	0.216	0.209	0.029	0.059	0.046	0.148	0.005	0.042

- We take $\tilde{f}_c \in FC$ to FF as a feature candidate with the highest candidate weight $W^{\mathcal{RI}}_{\tilde{f}_c,FF}$ - in this step we add v_4.

$$FF := \{v_5, v_4\}\,.$$

- We update FL set by excluding the selected feature $\tilde{f}_c$ from FC, i.e.

$$FL := \{v_1, v_2, v_3, v_6, v_7, v_8, v_9, v_{10}, v_{11}, v_{12}\}\,.$$

2^{nd} selection

- We choose feature candidates into *feature candidates set* FC:

$$FC := \{v_1, v_2, v_3, v_7, v_8, v_{10}, v_{11}, v_{12}\}$$

We removed v_6 and v_9, because $\mathcal{RI}_{4,6}^{\mathcal{AIC}} < 0$ and $\mathcal{RI}_{4,9}^{\mathcal{AIC}} = 0$.

- We check the weight of each feature candidate $f_c \in FC$:

$W_{f_c,\{v_5,v_4\}}^{\mathcal{RI}}$	v_1	v_2	v_3	v_7	v_8	v_{10}	v_{11}	v_{12}
	0.175	0.213	0.219	0.153	0.094	0.170	0.035	0.214

- We take $\tilde{f}_c \in FC$ to FF as a feature candidate with the highest candidate weight $W_{\tilde{f}_c,FF}^{\mathcal{RI}}$ - in this step we add v_3.

$$FF := \{v_5, v_4, v_3\}.$$

- We update FL set:

$$FL := \{v_1, v_2, v_7, v_8, v_{10}, v_{11}, v_{12}\}.$$

3^{rd} selection

- We choose feature candidates into *feature candidates set* FC:

$$FC := \{v_7, v_8, v_{10}, v_{11}, v_{12}\}$$

- We check the weight of each feature candidate $f_c \in FC$:

$W_{f_c,\{v_5,v_4,v_3\}}^{\mathcal{RI}}$	v_7	v_8	v_{10}	v_{11}	v_{12}
	0.230	0.150	0.338	0.091	0.452

- We take $\tilde{f}_c \in FC$ to FF as a feature candidate with the highest candidate weight $W_{\tilde{f}_c,FF}^{\mathcal{RI}}$ - in this step we add v_{12}.

$$FF := \{v_5, v_4, v_3, v_{12}\}.$$

- We update FL set:

$$FL := \{v_7, v_8, v_{10}, v_{11}\}.$$

4^{th} selection

- We choose feature candidates into *feature candidates set* FC:

$$FC := \{v_7, v_8, v_{10}, v_{11}\}$$

- We check the weight of each feature candidate $f_c \in FC$:

$W^{\mathcal{RI}}_{f_c,\{v_5,v_4,v_3,v_{12}\}}$	v_7	v_8	v_{10}	v_{11}
	0.304	0.219	0.547	0.128

- We take $\tilde{f}_c \in FC$ to FF as a feature candidate with the highest candidate weight $W^{\mathcal{RI}}_{\tilde{f}_c,FF}$ - in this step we add v_{10}.

$$FF := \{v_5, v_4, v_3, v_{12}, v_{10}\}.$$

- We update FL set:

$$FL := \{v_7, v_8, v_{11}\}.$$

Following the algorithm we finally receive

$$FF = \{v_5, v_4, v_3, v_{12}, v_{10}, v_7, v_8, v_{11}\}.$$

Now let us develop several models:

1. Full model without filtering features according to $\mathcal{AIC}$ - leukemia.exp.full: $Y\tilde{}g48 + g49 + g50 + g65 + g92 + g98 + g112 + g133 + g134 + g136 + g139$,
2. Model leukemia.exp.full with backward step procedure - leukemia.exp.full.back: $Y\tilde{}g50 + g65 + g134 + g139$,
3. Full model based on the features selected by the $\mathcal{AIC}$ algorithm - leukemia.exp.AIC.full: $Y\tilde{}g88 + g65 + g50 + g139 + g134 + g98 + g112 + g136$,
4. Model leukemia.exp.AIC.full with backward step procedure - leukemia.exp.AIC.full.back: $Y\tilde{}g88 + g65 + g50 + g139 + g134$.

The results obtained for these four models are summarized in Tab. 5.

model	N. of variables	N. of significant variables $\alpha = 0.1$	N. of significant variables $\alpha = 0.05$	AIC
leukemia.exp.full	11	2 (g134, g139)	2	46.015
leukemia.exp.full.back	4	4	3 (g50, g134, g139)	36.250
leukemia.exp.AIC.full	8	0	0	31.888
leukemia.exp.AIC.full.back	5	5	3 (g88, g50, g134)	28.682

We can see (Tab.5) that the $\mathcal{AIC}$ feature selection improved the model's $\mathcal{AIC}$. The full model with $\mathcal{AIC}$ selection (i.e. Model 3) has even better $\mathcal{AIC}$

than the first model with backward selection (i.e. Model 2). The backward selection applied to the third model has not only still improved the $\mathcal{AIC}$ index and reduced the dimensionality but has also made the coefficients significant.

6 Conclusions

A method suggested in this paper was prepared as an alternative to traditional selection methods especially for situations with multidimensional data. The most difficult problem we have to face in the project is a possible conflict between $\mathcal{AIC}$ minimization and improving significance of models' coefficients (the best models may appear as models with insignificant coefficients), since such model is completely useless for prediction. Both criteria are not independent and in some situations may lead to opposite conclusions: F-statistic value for testing assessment on two models (full model versus reduced) can be low which means that the reduced model is superior to the full one, while - in the other hand - $\mathcal{AIC}_{full} - \mathcal{AIC}_{reduced}$ is below zero which indicates that the full model is superior ([3], pp. 27-28). This undesired paradox definitely needs to be elaborated.

Acknowledgments This contribution is supported by the Foundation for Polish Science under International PhD Projects in Intelligent Computing. Project financed from The European Union within the Innovative Economy Operational Programme 2007-2013 and European Regional Development Fund.

References

[1] Akaike H., *Information theory and an extension of the maximum likelihood principle*, In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium Information Theory, Akademia Kiado, Bubapest, pp. 267-281, 1973.

[2] Bengtsson T., Cavanaugh J.E., *An improved Akaike information criterion for state-space model selection*, Computational Statistics and Data Analysis 50 (2006), 2635-2654.

[3] Bonate P.L., *Pharmacokinetic-pharmacodynamic modeling and simulation*, Springer, 2005.

[4] Burnham K.P., A.R., *Model Selection and Inference. A practical Information-Theoretic Approach*, Springer-Verlag, New York, 1998.

[5] Cover T.M, Thomas J.A, *Elements of Information Theory*, Springer, 2001.

[6] Hafidi B., Mkhadri A., *A corrected Akaike criterion based on Kullbacks symmetric divergence: applications in time series, multiple and multivariate regression*, Computational Statistics and Data Analysis 50 (2006), 1524–1550.

[7] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer, 2002.

[8] Lukacs P.M., Burnham K.P., Anderson D.R, *Model Selection Bias and Freedman's Paradox*, Springer, 2009.

[9] Shang J., Cavanaugh J.E., *Bootstrap variants of the Akaike information criterion for mixed model selection*, Computational Statistics and Data Analysis 52 (2008), 2004–2021.

[10] Zhao Z., Zhang Y., Liao H., *Design of ensemble neural network using the Akaike information criterion*, Engineering Applications of Artificial Intelligence 21 (2008), 1182–1188.