ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Dorota Rozmus**

# METHODS OF CLASSIFICATION ERROR DECOMPOSITIONS AND THEIR PROPERTIES

**Abstract.** The idea of error decomposition originates in regression where squared loss function is applied. More recently, several authors have proposed corresponding decompositions for classification problem, where 0-1 loss is used. The paper presents the analysis of some properties of recently developed decompositions for 0-1 loss.

**Key words:** classification error, prediction error, error decomposition.

## 1. INTRODUCTION

The main aim of aggregating models is to decrease prediction error. There are two ways of building such models: we can get new training subsets from the original set by selecting features or by selecting objects. A learning algorithm builds a model on the base of each subset and then single models are aggregated[1] into one model (see: Fig. 1).

The error of the aggregated model, given in formula (1), can be decomposed into three components that show how such factors, e.g.: the number of single models, the way of getting training subsets, the number of objects in subsets or parameters of single models influence the value of the error

$$Err_{arg} = E_Z\{E_y[L(y, \hat{y})]\} \tag{1}$$

where: $L(y, \hat{y})$ – a loss function.

* M.Sc., Departament of Statistics, Karol Adamiecki University of Economics, Katowice.

[1] A learning algorithm is a function which takes as sole a learning sample and outputs a classification or regression model.

The subscript $y$ in equation (1) denotes that the expectation is taken with respect to noise, and the subscript $Z$ – that the expectation is taken with respect to the predictions produced by single models, built by learning algorithm on the training subsets from set $Z$.

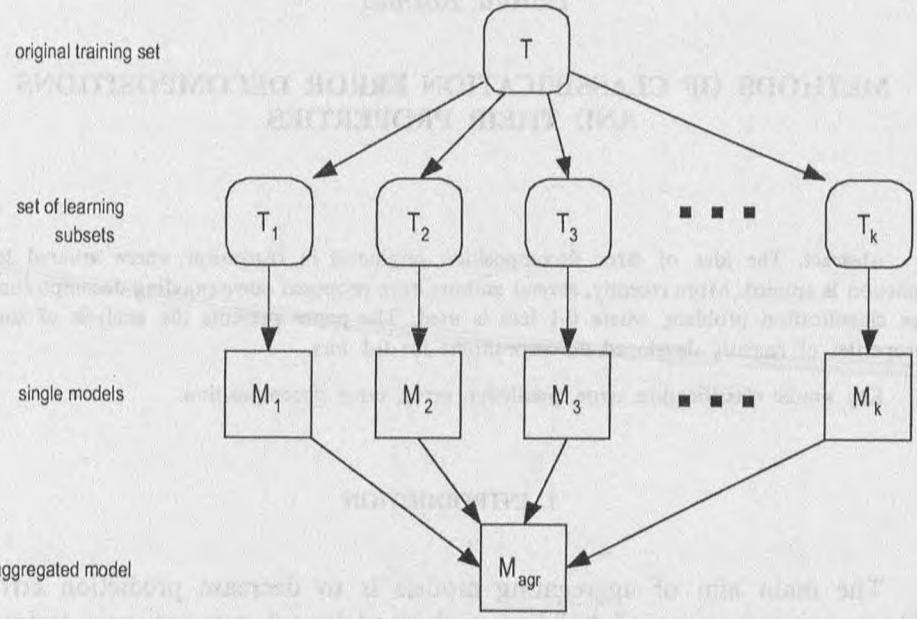


Fig. 1. Aggregated model

Source: own research.

## 2. PREDICTION ERROR DECOMPOSITION

The idea of error decomposition originates in regression, where squared loss function is applied, and it breaks down the error into three terms: noise ($N(\mathbf{x})$), bias ($B(\mathbf{x})$) and variance ($\mathrm{D}^2(\mathbf{x})$) (Geman *et al.* 1992)

$$Err_{agr} = E_Z\{E_y[(y - \hat{y})^2]\} = N(\mathbf{x}) + B(\mathbf{x}) + \mathrm{D}^2(\mathbf{x}) \tag{2}$$

Before defining the decomposition components, we should first introduce the idea of aggregated model prediction and optimal model prediction (Domingos 2000).

In general, the dependent variable $y$ is a nondeterministic function of predictors $\mathbf{x}$, so it is almost impossible to get the value of loss function

equal 0. But for a determined problem and determined loss function we may define the best possible model that ensures the optimal prediction of $y$

$$y_* = \arg\min_{y'} E_y[L(y, y')] \tag{3}$$

Prediction on the base of optimal model, for determined example $\mathbf{x}$, is such value of $y$ that minimises the expected value of loss

$$E_y[L(y, y_*)] \tag{4}$$

In the case of regression such prediction equals the expected value of $y$

$$y_* = E_y[y] \tag{5}$$

Prediction on the base of aggregated model, for any loss function and collection of training subsets in $Z$, is the prediction that minimises the expected value of loss

$$y_m = \arg\min_{\hat{y}'} E_Z[L(\hat{y}, \hat{y}')] \tag{6}$$

Aggregated model predicts such value $\hat{y}'$ whose average loss, relative to all the predictions obtained by means of single models, is minimum. In regression it is the mean of all single model predictions

$$y_m = E_Z[\hat{y}] \tag{7}$$

The noise of the learning algorithm on an example $\mathbf{x}$ in regression is defined in formula (8) and it is the loss coming from the difference between the real value of dependent variable and the value obtained on the base of optimal model

$$N(\mathbf{x}) = E_y[(y - y_*)^2] \tag{8}$$

It is an unavoidable component of the loss, incurred independently of the model. It is the hypothetical lower boundary of the error.

The bias of the learning algorithm on an example $\mathbf{x}$, using squared loss function, is systematic loss incurred by the value of dependent variable obtained on the base of optimal model relative to the prediction of aggregated model

$$B(\mathbf{x}) = (y_* - y_m)^2 \tag{9}$$

The variance of a learning algorithm on example x is the average loss incurred by single predictions relative to prediction on the base of aggregated model

$$D^2(\mathbf{x}) = E_Z[(y_m - \hat{y})^2] \tag{10}$$

The sum of the three elements is equal to the value of prediction error, when the squared loss function is applied (Geman *et al.* 1992).

## 3. CLASSIFICATION ERROR DECOMPOSITION

In recent years, several authors have tried to apply the idea of error decomposition to classification problems, where 0-1 loss is used. Then, the three components can be described as below.

Optimal model in classification associates each input x with the most likely class, according to the conditional class probability distribution

$$y_* = \arg\max_y P_y(y \mid \mathbf{x}) \tag{11}$$

The prediction of aggregated model in classification is the class receiving the majority of votes among all classes predicted by single models

$$y_m = \arg\max_{\hat{y}} P_y(\hat{y} \mid \mathbf{x}) \tag{12}$$

Noise in classification is equal 1 minus the probability of optimal classification

$$N(\mathbf{x}) = 1 - P_y(y_* \mid \mathbf{x}) \tag{13}$$

In regression bias was defined as difference between the value of dependent variable predicted by optimal model relative to aggregated model; in classification bias is given by indicator function

$$B(\mathbf{x}) = 1(y_* \neq y_m \mid \mathbf{x}) \tag{14}$$

So in classification biased observations are those, for which the classification on the base of aggregated model is different relative to classification of the optimal model.

The variance of a learner on example x in classification is equal 1 minus probability of the same classification as obtained by means of aggregated model among all classification by single models

$$D^2(x) = 1 - P_Z(y_m|x) \tag{15}$$

But it appears that in classification, by analogy to regression, the sum of such defined terms is not equal to the value of classification error (Domingos 2000):

$$Err_{agr} = E_Z\{E_y[1(y \neq \hat{y}|x)]\} \neq N(x) + B(x) + D^2(x) \tag{16}$$

More recently, several authors have proposed corresponding decompositions in classification problem. The difference in definitions of bias and variance can be attributed to disagreement over the properties that those terms should fulfil in the case of 0-1 loss. The most often used decompositions were proposed by: E. B. Kong and T. G. Dietterich (1995), R. Kohavi and D. H. Wolpert (1995), R. Tibshirani (1996), P. Domingos (2000) and L. Breiman (1996, 2000).

As the values of bias and variance are different depending on which decomposition was used, the article discusses how the values of those terms depend on different proposed definitions, and it tries to find concepts giving similar or even the same values.

Table 1 shows the main properties of bias and variance depending on concept of decomposition:

Table 1

Main properties of bias and variance, depending on concept of decomposition

| Authors of decomposition | Bias properties | Variance properties |
|---|---|---|
| Kong and Dietterich | it may be negative | it may be negative |
| Tibshirani | it may be negative | it may be negative |
| Kohavi and Wolpert | it is nonnegative | it is nonnegative |
| Breiman I | it is nonnegative | it is nonnegative |
| Breiman II | it is nonnegative | it is nonnegative |
| Domingos | it may be negative | it may be negative |

Source: own research.

## 4. BIAS, VARIANCE VALUES AND NUMBER OF PREDICTORS SELECTED FOR BUILDING A SINGLE MODEL

A benchmark dataset "Satimage", where there are 4435 objects (Blake *et al.* 1998), was chosen for experiments. For the training set there is a separate test set with 2000 observations. The objects in this set are fragments ($3 \times 3$ pixels) of the Earth area. The whole area has $82 \times 100$ pixels and each pixel represents the area of $80 \times 80$ meters. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the $3 \times 3$ neighbourhood. Each object belongs to one from six possible classes that denote the way of soil utilisation.

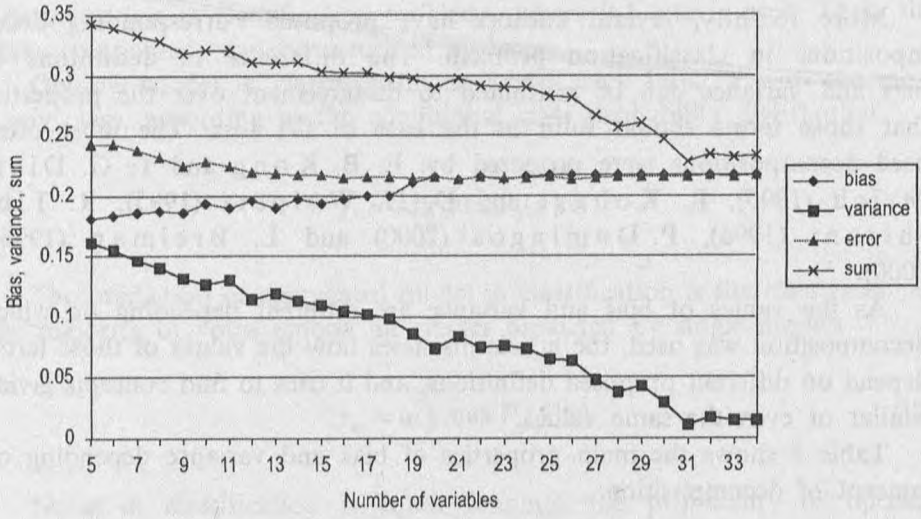


Fig. 2. Bias, variance, classification error according to number of variables randomly selected for building single models

Source: own research.

The aim of the first experiment is to verify how the value of bias and variance, depending on different concepts of calculating, will be formed according to the number of variables randomly selected for building single models. The number of models in aggregated model was stated as constant equal 100. Building of the single models started with 5 randomly selected variables and procedure was continued up to the moment, when 34 variables were taken, adding each time one more variable. Figure 2 shows what are the values of bias and variance calculated from general definition (formula

c(14) and (15)). It is clearly seen that the sum of these two terms is not equal to the value of classification error. Initially this sum is much higher than the error, and later it slowly comes more and more equal to it.

Thanks to Fig. 3, showing what the formation of bias calculated by means of different definitions is, it can be said that, generally, they indicate a very similar, increasing tendency. Initially significant differences in the values, according to the way of counting, come less and less apparent along with growth of number of features.

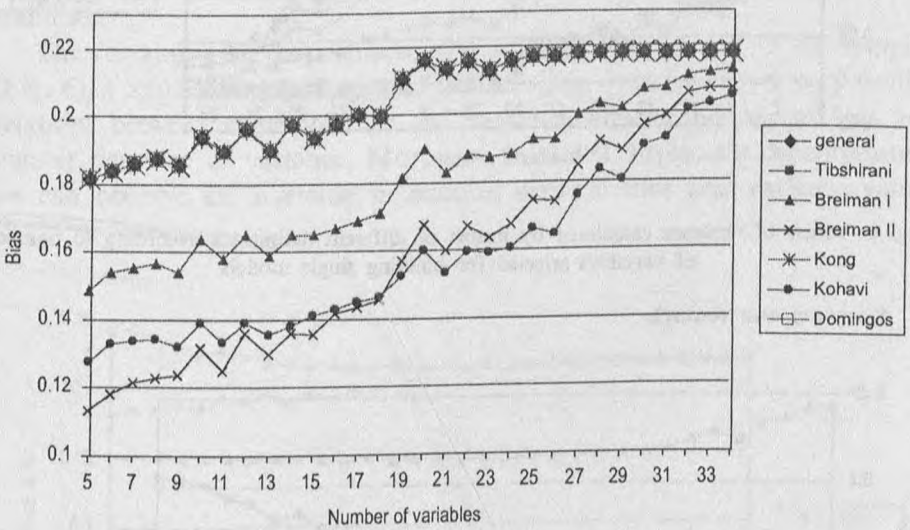


Fig. 3. Values of bias calculated by means of different definitions according to number of variables selected for building single models

Source: own research.

It is worth saying that the three concepts – by R. Tibshirani, P. Domingos, E. B. Kong and T. G. Dietterich – give the same values of bias, which additionally are identical as bias calculated from the general definition. All remaining concepts give lower values.

Taking the values of variance, calculated by means of different proposed definitions into consideration it can be said that, generally, they also show very similar tendency, but this time the trend is decreasing (Fig. 4). As it was for bias, differences in values are less and less significant.

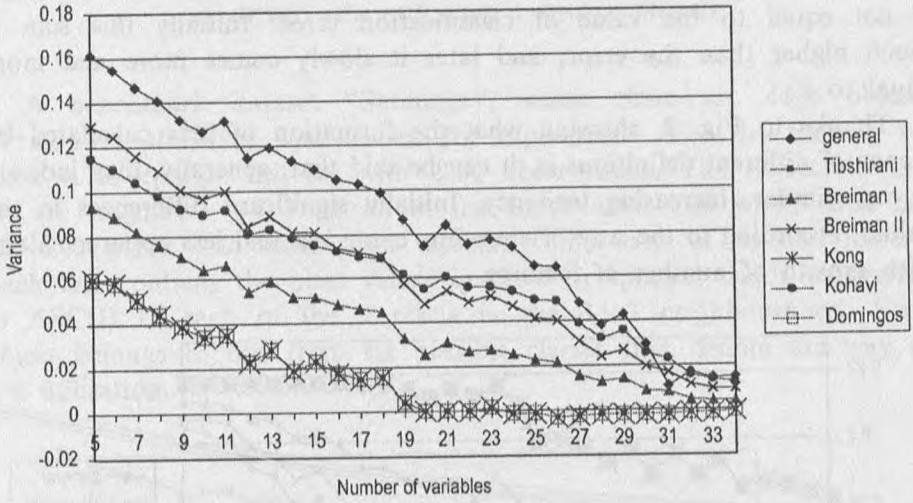


Fig. 4. Values of variance calculated by means of different definitions according to number of variables selected for building single models

Source: own research.

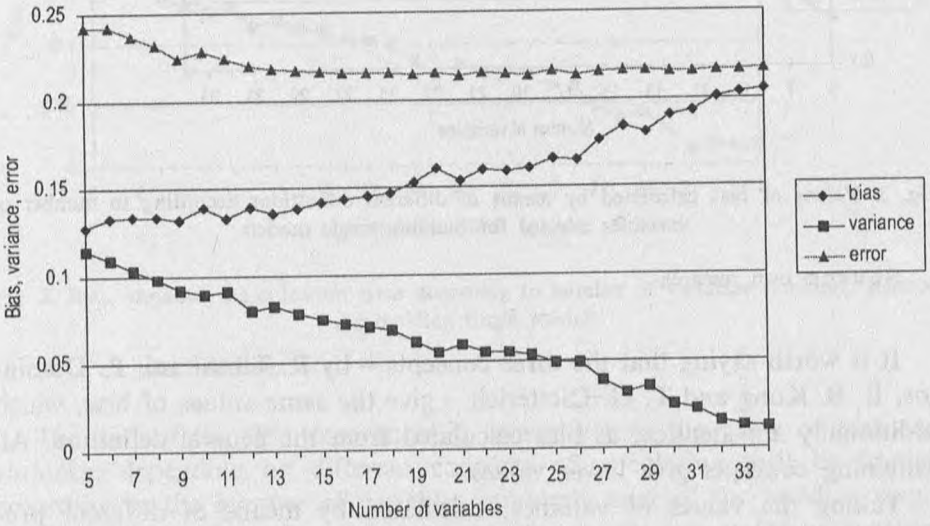


Fig. 5. Bias, variance, classification error according to definitions of R. Tibshirani, P. Domingos, E. B. Kong and T. G. Dietterich

Source: own research.

Definitions of R. Tibshirani, P. Domingos, E. B. Kong and T. G. Dietterich lead to the same, even negative values of variance, which form far away from the variance calculated from the general definition, that gives definitely the highest value.

It is also worth analysing the relations between bias and variance in single decompositions because they form differently. In some concepts, as e.g. in decompositions by R. Tibshirani, P. Domingos, E. B. Kong and T. G. Dietterich (Fig. 5) there are very low values of variance and the trend stabilises from some point. Together with it, there are very high, almost equal to the classification error, values of bias whose trend also stabilises.

The remaining decompositions, that is by R. Kohavi and D. H. Wolpert (Fig. 6), I and II Breiman's decomposition (Fig. 7 and 8) show very similar relations between values of bias and variance: regular increase of bias and regular decrease of variance. Moreover, in the II Breiman's decomposition we can observe an inversion in relation between bias and variance values (cross of lines).

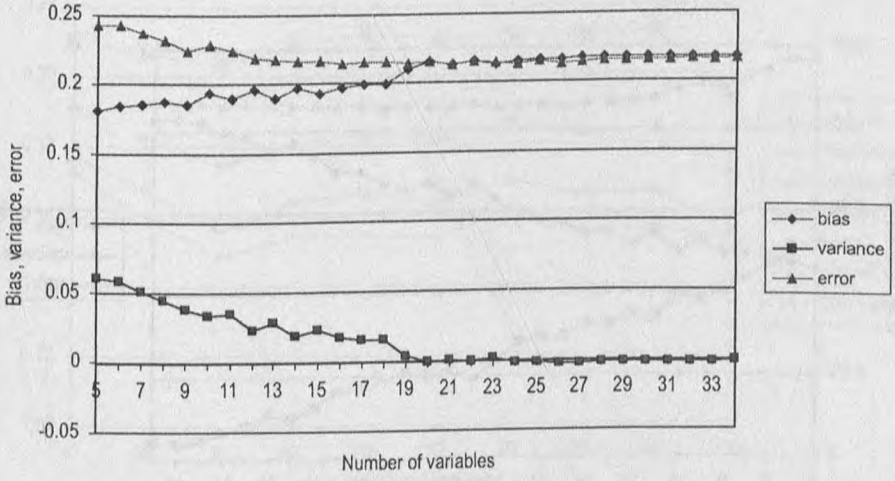


Fig. 6. Bias, variance, classification error according to first decomposition of R. Kohavi and D. H. Wolpert

Source: own research.

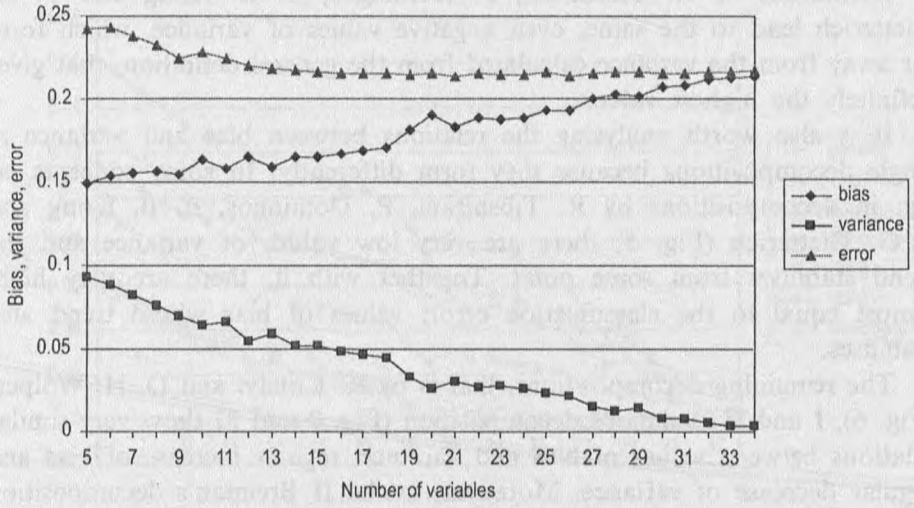


Fig. 7. Bias, variance, classification error according to I decomposition of L. Breiman

Source: own research.

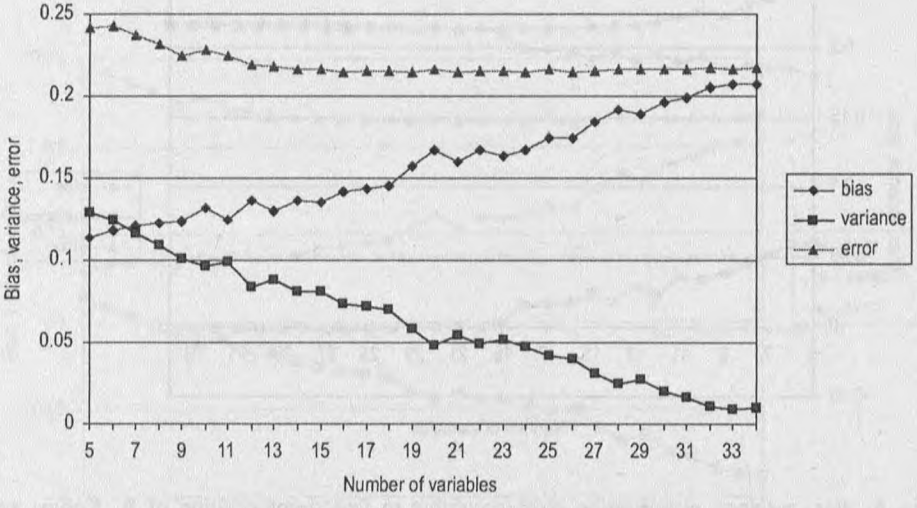


Fig. 8. Bias, variance, prediction error according to II decomposition of L. Breiman

Source: own research.

## 5. BIAS, VARIANCE VALUES AND NUMBER OF SINGLE MODELS IN AGGREGATED MODEL

The aim of the second experiment is to verify how will the values of bias and variance, calculated by means of different decompositions, form according to the number of single models in the aggregated one. In experiment building of an aggregated model started with 10 single models, and the procedure was continued up to 100, adding each time 10 more models.

Again, the sum of bias and variance, calculated from general definition, is not equal to the value of classification error.

Although it is rather difficult to notice a clear increasing or decreasing trend in values of bias, but it is seen that all proposed definitions show a very similar way of forming (Fig. 9). And again three decompositions – by R. Tibshirani, P. Domingos, E. B. Kong and T. G. Dietterich – give the same value, which is equal to the value obtained from the general definition.

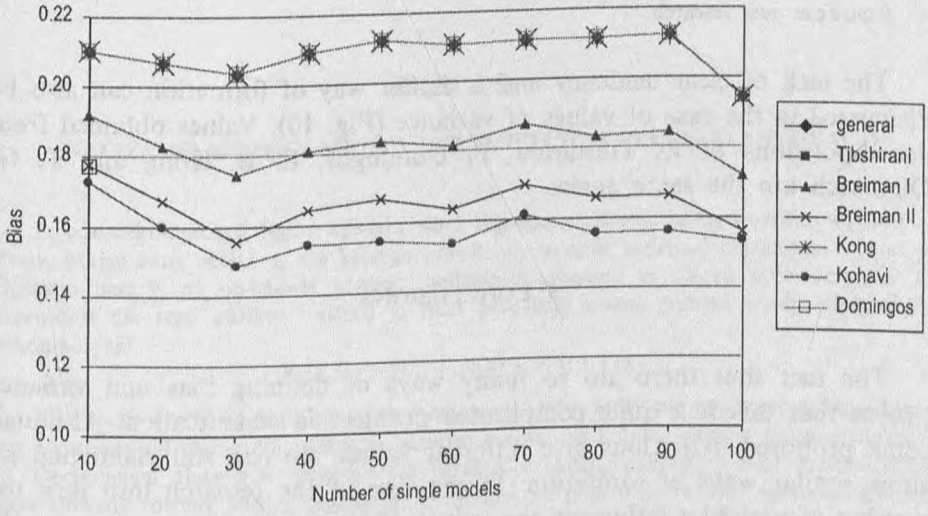


Fig. 9. Values of bias calculated by means of different definitions according to number of single models in aggregated model

Source: own research.

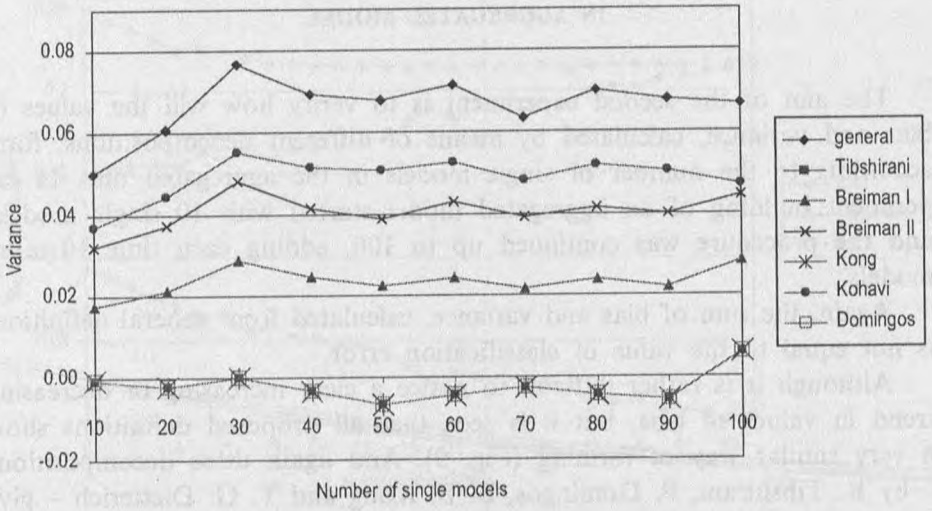


Fig. 10. Values of variance calculated by means of different definitions according to number of single models in aggregated model

Source: own research.

The lack of clear tendency and a similar way of formation can also be disscussed in the case of values of variance (Fig. 10). Values obtained from decompositions of R. Tibshirani, P. Domingos, E. B. Kong and T. G. Dietterich are the same again.

## 6. CONCLUSIONS

The fact that there are so many ways of defining bias and variance proves that this is a quite complicated problem in classification. Although some proposed definitions give different values, we can still claim that all show similar ways of formation. In the case of the research into how the number of variables influences the values of bias and variance, we can talk about a clear increasing trend of bias and decreasing trend of variance. In the experiment which aim was to verify how the number of single models influence the bias and variance we can say that even though there is no clear tendency, there are similarities in the way of formation.

Choosing one of the proposed ways of defining bias and variance in classification, it should also be noticed that different decompositions form the relations between bias and variance in a different way.

## REFERENCES

Blake C., Keogh E., Merz C. J. (1998), *UCI Repository of Machine Learning Databases*. Department of Information and Computer Science, University of California, Irvine.

Breiman L. (1996), *Arcing classifiers. Technical Report*, Department of Statistics, University of California, California.

Breiman L. (2000), *Randomizing outputs to increase prediction accuracy*, "Machine Learning", **40**, 3, 229–242.

Domingos P. (2000), *A unified bias-variance decomposition for zero-one and squared loss*, "Proceedings of the Seventeenth National Conference on Arificial Intelligence", AAAI Press, Austin, 564–569.

Geman S., Bienenstock E., Doursat R. (1992), *Neural networks and the bias/variance dilemma*, "Neural Computation", **4**, 1–58.

Kohavi R., Wolpert D. H. (1995), *Bias plus variance decomposition for zero-one loss functions*, "Proceedings of the Twelfth International Conference on Machine Learnig", Morgan Kaufmann, Tahoe City, 275–283.

Kong E. B., Dietterich T. G. (1995), *Error-correcting output coding corrects bias and variance*, „Proceedings of the Thirteenth International Conference on Machine Learnig", Morgan Kaufmann, 313–321.

Tibshirani R. (1996), *Bias, variance and prediction error for classification rules*, *Technical Report*, Department of Statistics, University of Toronto, Toronto.

*Dorota Rozmus*

## ANALIZA WŁASNOŚCI METOD DEKOMPOZYCJI BŁĘDU KLASYFIKACJI

Pojęcie dekompozycji błędu wywodzi się z regresji, gdzie stosuje się kwadratową funkcję straty. Mając dany obiekt $\mathbf{x}$, dla którego prawdziwa wartość zmiennej objaśnianej wynosi $y$, algorytm uczący, na podstawie każdego podzbioru uczącego ze zbioru prób uczących $Z$, przewiduje dla tego obiektu wartość $\hat{y}$. Błąd predykcji można poddać wtedy następującej dekompozycji:

$$E_Z\{E_y[(y-\hat{y})^2]\} = N(\mathbf{x}) + B(\mathbf{x}) + V(\mathbf{x}).$$

Błąd resztowy ($N(\mathbf{x})$) jest elementem składowym błędu, który nie podlega redukcji i który jest niezależny od algorytmu uczącego. Stanowi hipotetyczną dolną granicę błędu predykcji.

Obciążeniem algorytmu uczącego dla obiektu $\mathbf{x}$ ($B(\mathbf{x})$), nazywamy błąd systematyczny spowodowany różnicą między predykcją, otrzymaną na podstawie modelu optymalnego ($y_*$), a predykcją na podstawie modelu zagregowanego ($y_m$), gdzie $y_*$ i $y_m$ definiowane są jako

$$y_* = E_y[y], \quad y_m = E_Z[\hat{y}].$$

Wariancja dla obiektu $\mathbf{x}$ ($D^2(\mathbf{x})$) to przeciętny błąd wynikający z różnicy między predykcją na podstawie modelu zagregowanego ($y_m$) a predykcją uzyskaną na podstawie pojedynczych modeli ($\hat{y}$).

W literaturze pojawiły się także liczne koncepcje przeniesienia idei dekompozycji do zagadnienia klasyfikacji. Celem artykułu jest analiza własności różnych sposobów dekompozycji błędu przy zastosowaniu zero-jedynkowej funkcji straty.