

Dariusz AMPUŁA

Military Institute of Armament Technology (Wojtkowy Instytut Techniczny Uzbrojenia)

RANDOM FOREST IN THE TESTS OF SMALL CALIBER AMMUNITION

Losowy las w badaniach amunicji strzeleckiej

Abstract: *In the introduction of this article the method of building a random forest model is presented, which can be used for both classification and regression tasks. The process of designing the random forest module was characterized, paying attention to the classification tasks module, which was used to build the author's model. Based on the test results, a random forest model was designed for 7,62 mm ammunition with T-45 tracer projectile. Predictors were specified and values of stop parameters and process stop formulas were determined, on the basis of which a random forest module was built. An analysis of the resulting random forest model was made in terms of assessing its prediction and risk assessment. Finally, the designed random forest model has been refined by adding another 50 trees to the model. The enlarged random forest model occurred to be slightly stronger and it should be implemented.*

Keywords: random forest, predictor, node, feature, test

Streszczenie: *W artykule we wstępie przedstawiono metodę budowy modelu losowy las, którą można stosować zarówno do zadań klasyfikacyjnych, jak i do zadań regresyjnych. Scharakteryzowano proces projektowania modułu losowego lasu, zwracając uwagę na moduł zadań klasyfikacyjnych, który posłużył do budowy autorskiego modelu. Na podstawie posiadanych wyników badań, zaprojektowano model losowego lasu dla amunicji strzeleckiej kalibru 7,62 mm z pociskiem smugowym T-45. Wyszczególniono predyktory oraz określono wartości parametrów zatrzymania oraz formuły stopu procesu, na podstawie których zbudowano moduł losowego lasu. Dokonano analizy otrzymanego modelu losowego lasu pod kątem oceny jego trafności predykcji oraz oceny ryzyka. Na końcu, udoskonalono zaprojektowany model losowego lasu poprzez dodanie do modelu kolejnych 50 drzew. Powiększony model losowego lasu okazał się nieznacznie silniejszy i to on powinien być wdrożony do użytkowania.*

Słowa kluczowe: losowy las, predyktor, węzeł, cecha, badanie

1. Introduction

Random forest [4] is a set of simple trees (classification or regression) predicting the value of a dependent variable based on a set of independent variables (predictors). For classification tasks, the result is the predicted class membership about which the value of the qualitative dependent variable informs us. However, for regression tasks, the result is the predicted number.

Random forest according to Wikipedia [2] or random decision forest is a team method of machine learning for classification, regression and other tasks, which depend on building many decision trees during learning and generating a class, which is the dominant of classes (classification) or the predicted average (regression) of individual trees.

Random forests improve the tendency of decision trees to over-adapt to the training set. The first random forest algorithm was created by Tin Kam Ho using the random subspace method, which in the Ho formula is a way to implement the "stochastic discrimination" approach to the classification proposed by Eugene Kleinberg.

An extension of this algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forest" as a trademark. This extension combines the idea of "**bagging** - from **bootstrap aggregating**" introduced by Breiman and the random function selection presented by Ho and later independently developed by Amit and Geman to building a set of controlled variance decision trees.

The purpose of this article was to designing and building a "random forest" model based on the results of diagnostic laboratory tests of 7,62 mm ammunition with a T-45 tracer projectile [8]. The secondary purpose was to show that it is possible to create this type of model and its implementation for new diagnostic tests of small caliber ammunition but only at the stage of the evaluation module, i.e. having the results of the tests, we automatically evaluate them using the designed model and make a postdiagnostic decision.

2. The process of designing a random forest module

Random forest is a fully functional implementation of the algorithm developed by Breiman [1]. This approach can be used both for regression tasks (predicting quantitative values) and classification tasks (predicting the belonging of objects to specific classes).

A random forest consists of a number of simple trees that usually give much more accurate predictions than a single, even very complex decision tree. This prediction is obtained by the so-called voting for classification tasks or averaging for regression tasks.

Each of the component trees that create a random forest uses a random subset selected from all predictors. These subsets are independent, and each predictor can be used by many trees, i.e. a random selection of predictors for individual trees is made of so-called returning [6].

Random forests [5] must have decision trees for individual classifiers. More importantly, the forests are strongly based on the idea of joining into a single family possibly good and as weakly dependent as possible classifiers, which can, but does not have to be the basis for the operation of the so-called boosting. Breiman showed that the probability of making a classification mistake by a random forest increases as the appropriately defined correlation coefficient between trees increases and decreases as the so-called the strength of individual trees, that is (also properly defined) quality of classification carried out by a single tree.

Classification by a random forest takes place, as in the case of the bagging algorithm, i.e. a given observation vector is classified by all forest trees and finally classified into the class that obtained (ordinary, i.e. unweighted) majority of "votes".

Due to the fact that trees in random forests and in the bagging algorithm are learned on bootstrap pseudo – samples, and as a result about 1/3 of the elements of the original learning sample are not drawn to a specific pseudo-sample, for both families can give very simple methods for assessing the validity of attributes of the observation vector from point of view of the classification.

Implicitly software [7] divides the analyzed data results (objects) into a learning sample and a testing sample in a ratio of 70% to 30% and then builds a model based on the learning sample, and then assesses it on the basis of a separate testing sample. The program selects the best number of trees in the built model based on the accuracy of predictions in the testing sample.

In the classification model of a random forest presented in this article, the measure of the quality of this model is the frequency of incorrect classifications, which must be determined when designing the model.

One of the advantages of the "random forest" module is the fact that the predictions are determined by a set of decision trees, each of which uses a random subset of predictors, i.e. independent variables. This property is especially beneficial when we have a large number of predictors in the created model.

The "random forest" module also supports the possibility of missing data among predictors. If for some observation there is no data about the predictor used for a certain division, then this observation is assigned a value for the node subject to this division. In our case, the prepared database of results is complete and has no missing data.

When all formal requirements are introduced into the software during the design of our model, the classification model "random forest" is automatically built based on the analyzed data and parameters set by the author of the model.

3. Random forest build model for small caliber ammunition

During the design of the random forest model for the first diagnostic laboratory tests of the tested small caliber ammunition, a database of test results 7,62 mm ammunition with T-45 tracer projectile was prepared [3]. This is the so-called intermediate ammunition,

which contains in its construction the streak necessary to show the shooter the approximate path of the projectile to properly perform the imposed combat task. The so-called scientific – research inquiries, that are not authoritative to other research results, have been eliminated from the database. Tests carried out for the Ministry of the Interior was also not included.

For the building of the model, only tests were taken in which the type of test specified in the test methodology was one for test samples stored in warehouses of the Polish Army's economic departments, which means that only the tested lots of 7,62 mm ammunition with T-45 projectile stored in the storage subset specified in the test methodology as "K".

The purpose of all these restrictions was to create a homogeneous data set that will be the basis for building a random forest model for this ammunition.

When designing a random forest model for the 7,62 mm ammunition with T-45 projectile for the first laboratory diagnostic tests, 4 values of the tested features (predictors) were adopted, which were information obtained after the diagnostic tests, namely predictors: number of inconsistencies in the importance class A (LA), the number of inconsistencies in the importance class B (LB), the number of inconsistencies in the importance class C (LC) and the number of inconsistencies in the importance class D (LD).

The values of the predictors obtained during the tests were introduced into the random forest model and were the initial values on which this designed model was based. The values of these predictors were written in numerical form, i.e. if there were any inconsistencies during the test, the specific number of these inconsistencies was entered, if there were no inconsistencies in the given importance class, the value zero was entered. The exact characteristics of individual importance classes are presented in the test methodology [9].

During the building of this model, a number of necessary auxiliary parameters were introduced, which are a requirement for building a random forest model for the analyzed small caliber ammunition.

In summary, the subject of the analysis was a set of data obtained during the first diagnostic laboratory tests of 7,62 mm ammunition with T-45 projectile. The set of these data was randomly divided into a learning sample and a testing sample, which accounted for 30% of cases from the data set. Of course, you can opt out of the random division and specify an individual variable identifying our sample. In our case, however, we remained with the random division, thus trusting the directions indicated in the software. The learning sample was used to create the model, while the designed model was evaluated on the testing sample. Each analyzed of the lot of small caliber ammunition was characterized by the obtained results of the tested features of this ammunition, which were predictors in the random forest model being built.

4. The results of building random forest

To building of our random forest model, a classification model was adopted due to the fact that in our case there is a qualitative dependent variable marked "DEC", which means

obtained postdiagnostic decision after the first laboratory tests. The above decision may take the form of six different decisions: "B5", "B3", "BP", "Z", "PS" and "W". The exact description of possible diagnostic decisions is presented in the test methodology [9]. All data in the diagnostic database of 7,62 mm ammunition with T-45 projectile used during the design of this model were prepared according to the same key so that they form a certain homogeneity and integrity. During building of the random forest model, software [7] was used in which some auxiliary parameters were adopted to obtain optimum designed random forest model.

In order to give more weight to accurate prediction, in our case of classification, for selected classes, the term 'equal' was entered in the incorrect classification costs, which means that all elements of the matrix of incorrect classification costs, except for the main diagonal, will be given the value of one.

A priori probability was also assumed at the same level, i.e. the term "equal" was also entered. This parameter determines how likely it is, without any preliminary knowledge about the values of predictor variables in the built model, that a given case or object belongs to a given class. Therefore, we assume that the probabilities of classifying a given case or object into individual classes are the same for all classes of the dependent variable.

The next step in building of the random forest model for the analyzed test results was to determine the number of predictors, in our case these are 3 chosen by programme predictors and to set the number of trees (we set 100 by default), which means the condition for completing building of the model. By default, we also set the proportion of the random test sample to be 30% and the proportion for subgroups to be 50%. This value tells us what part of the learning sample outside the testing sample will be randomly selected for learning the individual trees that make up our random forest model. We set the minimum number of objects in the node to be divided to 10, which is one way to limit the size of the tree. We set the maximum number of levels as 10 and the minimum number of descendant we set as 5. Determining this size is helpful when component trees have end nodes that are too small. By default, we accept the maximum number of nodes as a value of 100. The latter value limits the number of nodes forming the component trees and if it is equal to this value, the building process is stopped.

In the parameters for stopping the learning process, we assume the number of cycles for determining the error value of 10. In the percentage error decrease field, enter the value 5, which gives the smallest decrease in error required after the execution of the previously specified number of cycles to determine the error to continue building the model. If this error drops by a smaller amount, the learning process is interrupted.

When designing a random forest model, you can see how the error in the testing and learning sample changed as the number of trees in the model increased. From fig. 1 it can be seen that the built model after reaching 15 trees has achieved some stabilization, which does not mean that subsequent trees do not contribute anything and this model should not be expanded. The program has finished creating the model on the number of 50 trees due to the fact that the error decreased by less than 5%, even if the number of trees entered previously into the random forest model was not reached. This figure shows that the

resulting random forest model was created based on 50 trees and all results were determined based on these 50 trees.

In the software [7] in the random forest module we can create predicted value sheets and calculate errors for the testing sample and the learning sample. An example of a fragment of the created sheet for the learning sample is shown in fig. 2.

The next step in the analysis of the random forest model built was to assess the accuracy of the prediction. The simplest tool used for this is the matrix of incorrect classifications.

The matrix of incorrect classifications obtained for our model for the learning sample is presented in fig. 3. This sheet presents the observed and predicted classifications and the probability of belonging to each class in the designed random forest model. This sheet shows, for example, that for the observed "PS" classes, the predicted error for the "BP" class is 3.23%. Then we define the risk assessment sheet for the learning and testing sample. These data are shown in fig. 4.

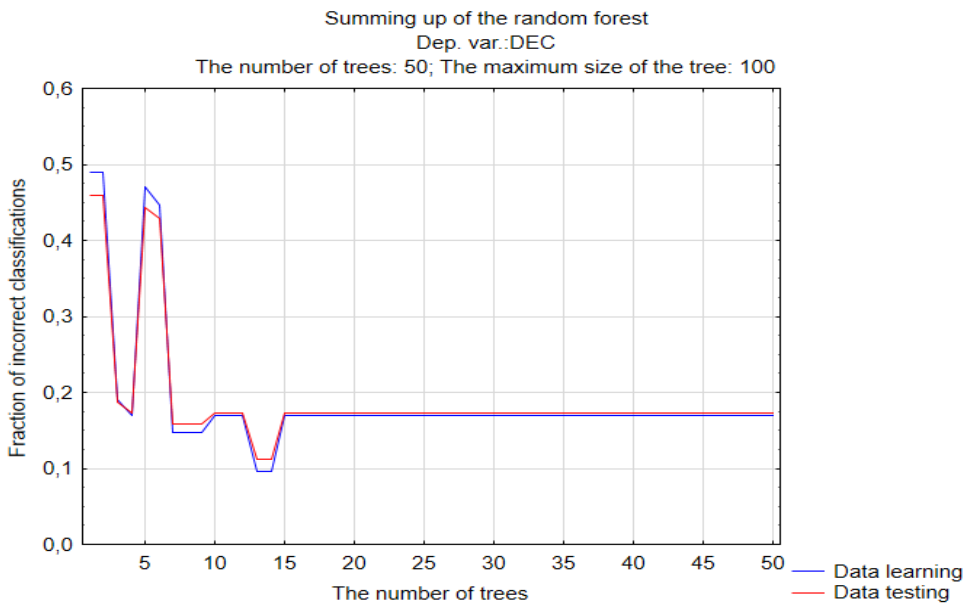


Fig. 1. The graph of changes fraction of incorrect classifications

Prediction (T-45 RB=1) Dependent variable: DEC Learning sample; The number of trees: 50								
Observed value	Predicted value	Probability B5	Probability PS	Probability BP	Probability Z	Probability B3	Probability W	
343	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
344	Z	Z	0,340000	0,000000	0,020000	0,640000	0,000000	0,000000
346	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
347	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
348	W	W	0,160000	0,000000	0,060000	0,000000	0,000000	0,780000
349	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
350	PS	PS	0,160000	0,800000	0,040000	0,000000	0,000000	0,000000
351	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
354	PS	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
355	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
357	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
358	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
359	Z	Z	0,340000	0,000000	0,020000	0,640000	0,000000	0,000000
361	B5	BP	0,380000	0,000000	0,620000	0,000000	0,000000	0,000000
363	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
364	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
365	Z	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
366	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
368	B5	B5	0,620000	0,000000	0,000000	0,020000	0,360000	0,000000
369	Z	Z	0,340000	0,000000	0,020000	0,640000	0,000000	0,000000

Fig. 2. Sheet of prediction values for learning sample – the printout of the application window

Matrix of classification (T-45 RB=1) Dependent variable: DEC Learning sample; The number of trees: 50								
	Observed	Predicted class B5	Predicted class PS	Predicted class BP	Predicted class Z	Predicted class B3	Predicted class W	Together in the line
Number	B5	166		29				195
% from column		98.22%	0.00%	55.77%	0.00%		0.00%	
% from line		85.13%	0.00%	14.87%	0.00%	0.00%	0.00%	
% from total		55.33%	0.00%	9.67%	0.00%	0.00%	0.00%	65.00%
Number	PS	1	27	1	1			31
% from column		0.59%	90.00%	1.92%	7.69%		2.78%	
% from line		3.23%	87.10%	3.23%	3.23%	0.00%	3.23%	
% from total		0.33%	9.00%	0.33%	0.33%	0.00%	0.33%	10.33%
Number	BP		2	11				13
% from column		0.00%	6.67%	21.15%	0.00%		0.00%	
% from line		0.00%	15.38%	84.62%	0.00%	0.00%	0.00%	
% from total		0.00%	0.67%	3.67%	0.00%	0.00%	0.00%	4.33%
Number	Z	1			10			11
% from column		0.59%	0.00%	0.00%	76.92%		0.00%	
% from line		9.09%	0.00%	0.00%	90.91%	0.00%	0.00%	
% from total		0.33%	0.00%	0.00%	3.33%	0.00%	0.00%	3.67%
Number	B3	1						1
% from column		0.59%	0.00%	0.00%	0.00%		0.00%	
% from line		100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	
% from total		0.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.33%
Number	W		1	11	2			35
% from column		0.00%	3.33%	21.15%	15.38%		97.22%	
% from line		0.00%	2.04%	22.45%	4.08%	0.00%	71.43%	
% from total		0.00%	0.33%	3.67%	0.67%	0.00%	11.67%	16.33%
Number	Total groups	169	30	52	13		36	300
% together		56.33%	10.00%	17.33%	4.33%	0.00%	12.00%	

Fig. 3. The matrix of incorrect classifications for learning sample – the printout of the application window

The risk in our classification case of a random forest was calculated as the fraction of cases classified incorrectly by the tree (for learning and testing sample), if different costs of incorrect classifications are set, the risk value is modified accordingly, i.e. expressed in relation to the overall cost. The standard error of assessment [1] is also given in the sheet. The calculated values of risk and standard error qualify the built model for use in practice.

The risk evaluation (T-45 RB=1) Dependent variable: DEC		
	Risk Evaluation	Standard error
Learning	0,170000	0,021687
Testing	0,172932	0,032793

Fig. 4. Sheet of values evaluation risk – the printout of the application window

The key results of the random forest model are the values of predictor importance, which indicate which variables most strongly affect the belonging to particular classes. As can be seen from the table in fig. 5, the most important predictor in our model is the LA class value. The importance of predictors in analyses for classification is calculated as the sum of the increases of node purity after all nodes of the tree and is expressed as a fraction of the maximum sum for all predictors.

Importance of predictors (T-45 RB=1) Dependent variable: DEC		
Variable rank	Importance	
LA	100	1,000000
LD	63	0,629570
LC	56	0,558724
LB	32	0,316941

Fig. 5. The ranking of importance predictors – the printout of the application window

As a result of the completion of the process of building a random forest model, we received a model consisting of 50 trees. The schema of the sample tree (No. 1) from our random forest is shown in fig. 6. From this figure it can be seen that the tree has 7 divide nodes and 8 end nodes. Each node contains the node ID, node size, selected dependent variable category, and histogram of dependent variables selected for the given node. Thus, in the obtained model we have 50 trees with different numbers of divide and end nodes. Using tree graphs or sheets with their structure, we can test individual trees making up the obtained random forest model. If, as a result of building your random forest model, you get a model with a large number of trees, then it may turn out that the analysis of all trees is impractical because it requires a lot of time.

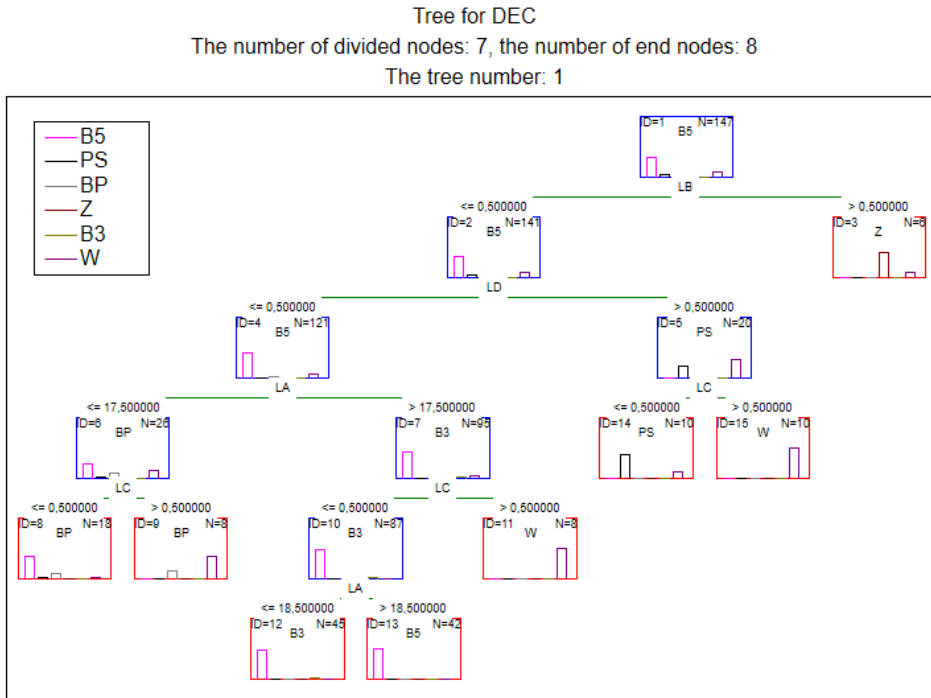


Fig. 6. The schema of the tree No.1 from random forest model

The detailed structure of tree number 1 from a random forest is shown in fig. 7. This table describes the data in all nodes of this tree, the number of individual node classes, the selected class in a given node, and the variable division criterion associated with the split constant value.

The structure of the tree (T-45 RB=1)												
Dependent variable: DEC												
Tree number: 1												
Descendant node 1	Descendant node 2	Size of the node	N class B5	N class PS	N class BP	N class Z	N class B3	N class W	Chosen class	Divide variable	Divide constant	
1	2	3	147	98	9	5	5	2	28	B5	LB	0,50000
2	4	5	141	98	9	5	0	2	27	B5	LD	0,50000
4	6	7	121	98	1	5	0	2	15	B5	LA	17,50000
6	8	9	26	13	1	5	0	0	7	BP	LC	0,50000
8			18	13	1	3	0	0	1	BP		
9			8	0	0	2	0	0	6	BP		
7	10	11	95	85	0	0	0	2	8	B3	LC	0,50000
10	12	13	87	85	0	0	0	2	0	B3	LA	18,50000
12			45	43	0	0	0	2	0	B3		
13			42	42	0	0	0	0	0	B5		
11			8	0	0	0	0	0	8	W		
5	14	15	20	0	8	0	0	0	12	PS	LC	0,50000
14			10	0	8	0	0	0	2	PS		
15			10	0	0	0	0	0	10	W		
3			6	0	0	0	5	0	1	Z		

Fig. 7. The structure of the tree No.1 from random forest model – the printout of the application window

The designed and built random forest model can be used for new test results for new tested lots of 7.62 mm ammunition with T-45 projectile. The corresponding code generated in the form of PMML can be used to implement this model. By running the module "Rapid implementation of predictive models" and opening the previously generated PMML code, we can predict postdiagnostic decisions for new predictor values.

After analyzing the built random forest model and concluding that it is not strong enough, it is possible to improve the existing model. This can be done by adding new trees giving the number of these additional trees. In our case, such an analysis was carried out and another 50 trees were introduced into the model. As a result of the process of building an enlarged random forest model, we received slightly lower values of the risk of assessment and standard error, as shown in Figure 8. In this model a smaller error is made when determining the fraction of cases incorrectly qualified by the tree for the teaching and testing sample.

The risk evaluation (T-45 RB=1)		
Dependent variable: DEC		
	Risk Evaluation	Standard error
Learning	0,146667	0,020425
Testing	0,157895	0,031618

Fig. 8. Sheet of values evaluation risk for enlarged model – the printout of the application window

The graph showing how the error in testing and learning sample changed as the number of trees in the model increased, in an enlarged random forest model is shown in fig. 9.

The built model after reaching 15 trees, as before, achieved some stabilization, then with 53 ÷ 54 trees there was a slight decrease in the fraction of incorrect classifications, which remained to 90 trees. Such a curve is confirmed by a decrease in the value of risk assessments and standard error for learning and testing sample.

In order to perform a full analysis of the designed random forest model, another 50 trees were introduced into this model, it was built and it was found that enlarging the model by another 50 trees did not contribute anything to this model, i.e. none of the parameters indicating a stronger new model not reached.

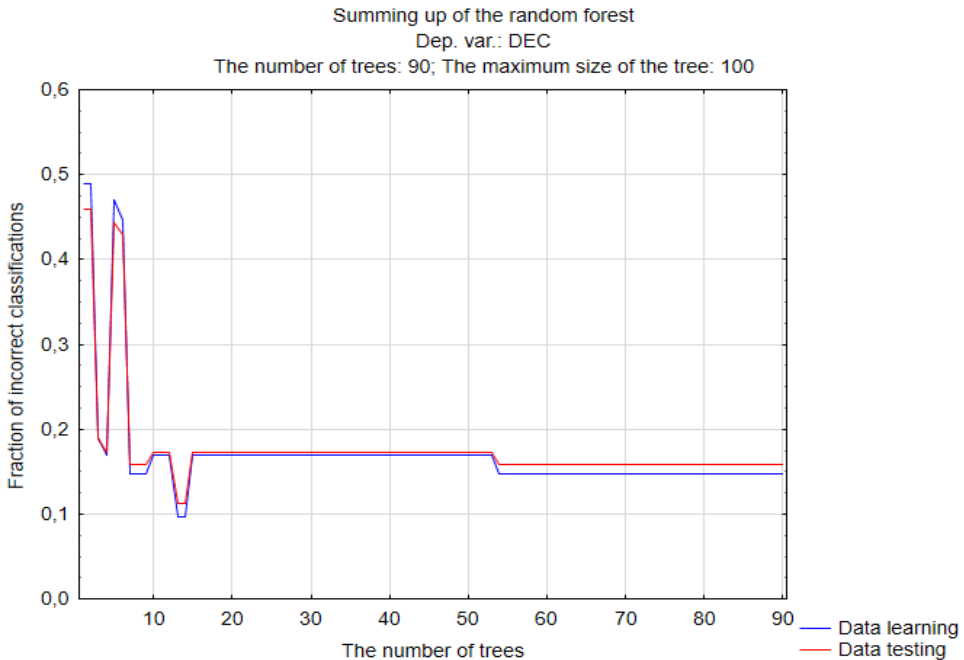


Fig. 9. The graph of changes fraction of incorrect classifications for enlarged model

The module for rapid implementation of predictive models of built models for new predictor values was tested for both the 50 trees model and the 90 trees model. This module calculated the value of the error made, calculated as the ratio of the error fraction to the average square error ex post, which was 0,170901 for the 50 trees model and 0,150115 for the 90 trees model. As you can see, the 90 trees model makes a smaller error when predicting the qualitative dependent variable.

It can therefore be said with certainty that the 90 trees model is slightly stronger and it should be implemented and used when determining the prediction for the new tested lots of 7,62 mm ammunition with T-45 projectile.

5. Summary

A random forest model for 7,62 mm ammunition with T-45 projectile was designed and built in the article. The diagnostic data base has been developed to meet the specific requirements of this method. The dual purpose defined at the beginning of the article has been fully achieved.

When designing and building our model of a random forest, due to the qualitative dependent variable, the necessary parameters were introduced that led to the development of optimum model. The resulting model of a random forest contains in its formulas all

possible analyzed classes, i.e. possible postdiagnostic decisions, as shown in the above figures.

Basically, "Random Forest" is a machine learning method in which we obtain the so-called "Black box" often giving accurate predictions, but without the possibility of easy interpretation. To perform a thorough analysis of the results of built random forest model, look for the so-called hyper – parameters at which the resulting model will be the strongest in its operation.

After analyzing the designed models of a random forest and based on the final results obtained from these models, a model consisting of 90 trees was adopted as the one with optimum performance parameters. This model is obviously not a perfect model, but it has satisfactory validity and is useful in operation. It is possible to design a better model of a random forest, however, this requires using more data results on which this model was based. Due to the fact that these new test results are not available, you should be content with those that are available.

The designed model of a random forest is after neural networks and decision trees, the third method of artificial intelligence, which can be successfully used to determine postdiagnostic decisions based on diagnostic laboratory tests. You can even building three assessment models for a specific type of ammunition being analyzed and check the correctness of obtaining the correct postdiagnostic decision using these three models. Which of these models is optimum - the answer to this question is difficult and can be obtained after conducting a number of additional studies in this topic. At the moment, each of the designed models can be used to evaluate the results of diagnostic tests, and in the opinion of the author of the article, each of them will work in the assessment procedure.

All these three methods work at a similar high level of probability of obtaining the correct value of the dependent variable sought by making predictions for the new lots of a particular type of ammunition. The elimination of the human factor from the assessment process, which may contribute to making assessment errors, is sufficient for the designed artificial intelligence assessment models to be implemented in diagnostic tests of all technical objects, which are elements of ammunition.

In this article, a random forest model was designed for 7,62 mm ammunition with T-45 projectile, but it is also possible to develop such models for other types of small caliber ammunition. The implementation of these new artificial intelligence methods for conducting diagnostic testing elements of ammunition is an indispensable necessity with which decision-makers in managerial positions in this test sector must also identify.

6. References

1. Breiman L., Friedman J., Olshen R.A., Stone C.J.: Classification and regression trees – 1984 r.
2. https://wikipedia.org/wiki/Las_losowy
3. Cards from laboratory tests of 7,62 mm ammunition with T-45 projectile – archive Military Institute of Armament Technology (MIAT).
4. Electronic handbook „Statistica” – Statsoft Poland 2020 r.

5. Koronacki J., Ćwik J.: Statystyczne systemy uczące się – Academic Publishing House EXIT, Warsaw 2008 r., pp. 162÷164.
6. Łapczyński M., Demski T.: Data mining – predictive methods – materials from course, Statsoft Poland 2019 r., p. 80÷87.
7. Statistics 13.3 PL – Statsoft Poland 2018 r. – computer software.
8. Collective work – Ammunition of land forces – Publishing House Ministry of National Defence, Warsaw 1985 r. – Uzbr. 2307/83, pp. 212÷220.
9. Collective work – Methodology of diagnostic tests of small caliber ammunition after long storage – Index N-5003a – 1986 r. – archive MIAT.

