Geosci. Model Dev. Discuss., 8, 5419–5435, 2015
www.geosci-model-dev-discuss.net/8/5419/2015/
doi:10.5194/gmdd-8-5419-2015

Geoscientific Model Development Discussions
Open Access

Application of all relevant feature selection for failure analysis of parameter-induced simulation crashes in climate models

W. Paja[1], M. Wrzesień[2], R. Niemiec[2], and W. R. Rudnicki[3,4]

[1]Department of Computer Science, Faculty of Mathematics and Natural Sciences, University of Rzeszów, Rzeszów, Poland
[2]Department of Artificial Intelligence and Expert Systems, Faculty of Applied Informatics, University of Information Technology and Management, Rzeszów, Poland
[3]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland
[4]Department of Bioinformatics, University of Białystok, Białystok, Poland

Received: 09 May 2015 – Accepted: 15 May 2015 – Published: 13 July 2015

Correspondence to: W. R. Rudnicki (w.rudnicki@icm.edu.pl)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

The climate models are extremely complex pieces of software. They reflect best knowledge on physical components of the climate, nevertheless, they contain several parameters, which are too weakly constrained by observations, and can potentially lead to a crash of simulation. Recently a study by Lucas et al. (2013) has shown that machine learning methods can be used for predicting which combinations of parameters can lead to crash of simulation, and hence which processes described by these parameters need refined analyses. In the current study we reanalyse the dataset used in this research using different methodology. We confirm the main conclusion of the original study concerning suitability of machine learning for prediction of crashes. We show, that only three of the eight parameters indicated in the original study as relevant for prediction of the crash are indeed strongly relevant, three other are relevant but redundant, and two are not relevant at all. We also show that the variance due to split of data between training and validation sets has large influence both on accuracy of predictions and relative importance of variables, hence only cross-validated approach can deliver robust prediction of performance and relevance of variables.

1 Introduction

Development of realistic models of climate is one of the most important areas of research due to dangers posed by global warming. It is by no means a trivial task since it involves parameterisation of many processes that are not directly solved within model. It has been shown by Lucas et al. (2013) that certain combinations of these parameters, lead to failure of a model, despite that each individual parameter has a reasonable value. Authors of this study performed 480 simulations using with systematically varied combinations of 18 parameters of the Parallel Ocean Program (POP2) (Smith et al., 2010) module in the Community Climate System Model Version 4 (CCSM4) (UCAR, 2010). They have determined by means of sensitivity analysis and machine learning al-

GMDD
8, 5419–5435, 2015

gorithms that 8 of these parameters contributed to the failure. They had applied support vector machine (SVM) (Vapnik, 2000) classification from machine learning to quantify and predict the probability of failure as a function of the values of 18 from POP2 parameters. The causes of the simulation failures were determined through a global sensitivity analysis. Combinations of 8 parameters related to ocean mixing and viscosity from three different POP2 parameterizations were then determined as the major sources of the failures. These 8 parameters were indicated as targets for more detailed research.

These results are somewhat disappointing, since the number of parameters is still rather high. Hence we decided to check whether more elaborate method for analysis could decrease this number further. We have observed potential weak points of the analysis performed by Lucas and co-workers, namely, they have not fully taken into account that the apparent importance of a variable for classification may be in fact result of a spurious fluctuation. The problem is most acute when a sample used for machine learning algorithm is small. In such a case random fluctuation may introduce spurious correlations within data, which can be utilized by classification algorithm for model building. The appropriate procedure should be applied to minimize influence of such random correlations on final results.

Lucas and co-workers have also analyzed impact of the decision variable that is used for classification on the quality of results. While the models were built as an ensemble of learners built on the bootstrap samples of the training set, the evaluation of classification performance was based on a single split of data between training set and test set. This setup was due to the construction of the study – simulations for the validation set were performed after the predictions have been made. While this is very honest method for verification of the predictions, however, it precludes estimation of statistical uncertainty of the result. In particular, it is impossible to say whether observed differences between classification accuracy observed for different decision functions are significant or do they arise due to statistical fluctuations.

Current study is devoted to the reanalysis of the data. It aims in minimizing influence of random fluctuations on the final results. Our aim was to establish all variables that

GMDD
8, 5419–5435, 2015

truly contribute to the final result of the simulations, i.e. whether the simulation was finished successfully or it crashed. To this end we use contrast variables that carry no information on decision variable, apply Boruta algorithm for all-relevant feature selection and extensive Monte Carlo sampling. We also compare quality of classification for several subsets of variables used for prediction of simulation result, to perform parallel check of relevance of variables.

2 Methods

Similarly to the original work, we rely on the machine learning algorithms to identify parameters that critically influence the fate of the simulation. Fundamental idea is that when classification algorithm can predict result of the algorithm, i.e. successful completion of simulation or crash, using only the information on the values of certain combinations of selected parameters, then these parameters are indeed responsible for the result. In the original paper authors performed true prediction and achieved good accuracy. In the current study we are limited to virtual prediction only, i.e. we can split the entire dataset into training and validation sets. We then build a model using training set and check it's quality performing virtual prediction on validation set and comparing the predicted results with the true ones. One can take advantage of virtualisation to obtain information about probability distribution of results. To this end one can perform multiple virtual experiments, with different splits between training and validation sets, and perform classification experiment on each of these splits. The results of individual trials will differ in most cases, allowing one to draw conclusions not only about mean values but also about variance and even shape of probability distribution.

In the original work the authors rely on an ensemble of SVM (Vapnik, 2000) learners, each member of the ensemble obtained on different subsample of the training set. The classifier was then used for prediction of the simulation result for the validation set. However, we use Random Forest (Breiman, 2001) as the classification algorithm,

GMDD
8, 5419–5435, 2015

and instead of sensitivity analysis we apply the all-relevant feature selection algorithm Boruta (Kursa and Rudnicki, 2010; Kursa et al., 2010).

Random Forest is en ensemble algorithm based on decision trees. To ensure low correlation between elementary learners, each tree is grown using different random subsample of the original data set, moreover, each split in the tree is built using only a random subset of the predictor variables. It is robust "of the shelf" algorithm that is easily applicable to various classification and regression tasks. It has only few control parameters and usually it does not need fine tuning for particular problem under scrutiny. In many cases it has performance comparable or even better than state of the art classifiers and it rarely fails. Big advantage of the algorithm is that it estimates by internal cross-validation both the estimate of the classification error and of the importance of variables. To estimate the latter it measures how much the accuracy of base learners is decreased when information about variable in question is removed from the system.

Boruta algorithm for all-relevant feature selection uses Random Forest importance measure to infer their relevance. To this end it extends the information system by variables that are non-informative by design – the so-called contrast variables. It then compares apparent importance of the original variables with that of the non-informative ones. It performs this multiple times using different realizations of the non-informative variables and performs statistical test. Algorithm finds both strongly and weakly relevant variables. The notions of strong and weak relevance were introduced by (Kohavi and John, 1997) in the context of ideal classification algorithm. Features are *strongly relevant* when removing them from description always results in decreased classification accuracy. Features are *weakly relevant*, when their removal in some cases may decrease classification accuracy. For more detailed discussion of relevance and Boruta algorithm see (Kohavi and John, 1997; Rudnicki et al., 2015). Algorithm has been used in different fields, including bioinformatics, remote sensing, bacteriology and medicine (Aagaard et al., 2012; Ackerman et al., 2013; Buday et al., 2013; Duro et al., 2012;

GMDD
8, 5419–5435, 2015

Herrera and Bazaga, 2013; Leutner et al., 2012; Ma et al., 2014; Menikarachchi et al., 2012; Saulnier et al., 2011; Strempel et al., 2013).

The climate simulations dataset is highly biased towards successful completion of simulation. Only 46 cases out of 540 are failures. Such unbalanced datasets are often difficult for classification, because automatic selection of the majority class results in good, but useless, classification accuracy. In such a case no information is gained and hence one cannot perform feature selection. In the current study this problem was avoided by constructing ten balanced subsamples of training set, and performing feature selection on each subsample. Each subsample consisted of all objects from minority class (failed simulations) and 1/11th of majority class (successful simulations). Procedure was repeated 60 times. Altogether all relevant feature selection was performed 660 times. In order to check specificity of the feature selection each dataset was extended by contrast variables. Each original variable was duplicated and it's values were randomly permuted between all objects. Hence a set of non-informative by design shadow variables was added to original variables. The number times when the shadow variables were selected as important gives estimate of the expected level of false discovery. The variables that were selected as important significantly more often than random, were examined further, using different test.

The second test probing the importance of variables was performed by analysing the influence of variables used for model building on the prediction quality. The first experiment revealed four variables that were classified as important by Boruta in all, or nearly all, of 660 trials. These variables were considered to form a core variable set, and model built using these variables was used as a reference. We examined whether removing one of the core variables and whether adding another variable respectively decreases or increases the classification quality measured by AUC. The extension of the core test was examined for three variables that were classified in the first test as important significantly more often than randomised variables.

The test was performed similarly to the one reported in the original study, with one important extension. The data set was randomly split into training set containing 360

GMDD
8, 5419–5435, 2015

objects and validation set containing 180 objects. The split was performed separately for minority and majority class, so the number of minority class objects in each training set was 32 and in the validation set it was 14. The procedure was repeated 30 times and results of 30 repetitions were analysed. In the original design only one split of the data between training set and validation set was reported, what severely influenced the results of the analysis.

3 Results and discussion

The summary of the results of the study is presented in the Table 1. The V1 and V2 variables were deemed important in all 660 cases. Variables V13 and V14 were deemed important in nearly all cases – 593 and 623 cases, respectively. All these variables were also indicated as most important by Lucas et al. (2013). However, the results don't agree so well for other variables. Lucas et al. indicated variables V4, V5, V16 and V17 as important but their influence on final result was much weaker than that of the first group. In the current study the variables V4 and V16 were deemed important by Boruta for 44 and 66 subsamples, respectively. In both cases the number is significantly higher than the average for random variables, which was obtained as 25 ± 9. On the other hand variables V5 and V17 were deemed important for 19 and 17 subsamples, respectively, and these numbers are lower than average for random variables. Moreover, variable V9, which was not indicated as important by Lucas et al. (2013), was deemed important for 62 subsamples.

Hence the first experiment confirmed importance of variables V1 and V2, has shown that importance of V13 and V14 is nearly universal, it also confirmed weak importance of variables V4 and V16. On the other hand the importance of variables V5 and V17 was not confirmed with our method, instead variable V9 was found to be weakly important. The example result of Boruta run for an interesting sample is presented in Fig. 1. In this sample the importance was confirmed for variables V9 and V16, whereas variable V13 was deemed irrelevant. The importance of V4 was higher than that of

GMDD
8, 5419–5435, 2015

highest random variable, but only barely so, and hence the final decision of Boruta was "tentative".

One should note, that Boruta is all-relevant feature selection algorithm that aims at finding both strongly and weakly relevant variables, as defined by Kohavi and John. The second test aimed at discerning between strongly and weakly relevant variables. In the case of V1, V2 removal of the variable from the core dataset resulted in dramatic drop of AUC, confirming that these variables are truly informative, see Table 1 and Fig. 2. In the case of V14 the difference in AUC – referenced further as Δ(AUC) – was smaller, but still statistically significant, whereas for the V13 the Δ(AUC) was much smaller than the standard deviation. Similarly, adding either of three remaining variables, namely V4, V9 and V16, to the core set, lead to increase of the AUC by insignificant amount, see Table 1 and Fig. 2. Another auxiliary metric that can be used to evaluate relevance of variables, is the number of samples in which AUC for model containing the variable is higher than that for model built without that variable. The results of this metric are consistent with results for the Δ(AUC) – it is 30 for both V1 and V2 and 26 for V14 and these are only results that are significantly different from random ones. Therefore one can conclude, that only three variables, namely V1, V2 and V14 are *strongly relevant*, whereas the remaining variables are *weakly relevant*.

One should note that the results of the second test were highly variable and largely dependent on the split of data between test and validation sets. It is illustrated in Fig. 3 and examples of the results from several samples are given in Table 2. The highest AUC obtained in the experiment was 0.990 for model built using core variables and V16 in sample #12. In the same sample the AUC for model built from core-V2 was 0.888. On the other hand for sample #1 the highest AUC was obtained for model built on core +V9 and it was 0.879. Also the relative importance of variables depends strongly on the test sample. For example adding of variable V4 to the core set can improve AUC by as much as 0.032 (sample #22) or decrease it by 0.006 (sample #6). Similarly for V16 AUC can decrease by 0.015 (sample #6) or increase by 0.016 (sample #22). Most interestingly, removing variable V13, which was deemed relevant by Boruta in nearly 90 %

GMDD
8, 5419–5435, 2015

of samples, can either decrease the AUC by 0.011 (sample #6) or increase it by 0.030 (sample #22). This results show that one cannot rely on a single split between training set and test set for estimate of influence of parameters, and that only the average over sufficiently large alternative split can give robust estimates.

The average of the cross-validated AUC obtained for three strongly important variables, namely V1, V2 and V13, was 0.924. The highest average AUC was obtained for model built using five variables, namely {V1, V2, V9, V13, V14}, nevertheless the value AUC = 0.931 was not significantly higher than the value obtained for simpler model built using only three variables. The small differences in AUC arise due to small improvements for assigning the probability of failure of the simulation. Such improvement results in small shift in the ranking from least probable to most probable to fail, without actually improving the error rate at the cost of including two more variables in the model.

4 Conclusions

Our reanalysis of the results of 540 simulations is in general qualitative agreement with the results of Lucas et al. The results of the simulation can be predicted with fairly good accuracy using machine learning approach, and two different methods give very close results. The cross-validated AUC reported by Lucas et al. by ensemble of SVM classifier was 0.93. In the current study the average of the cross-validated AUC obtained for three strongly important variables, was 0.924.

We have shown by cross-validation that the AUC reported for the prediction experiment performed by Lucas et al. falls within the range of values that can be expected in such prediction, however, one should not assign any weight to the particular value obtained. If the split between training set and test set was set differently the resulting AUC for prediction could be any number between 0.88 and 0.99.

Two most important conclusions for the climate modelling community are following. Firstly, the efforts on improving the numerical stability of simulations should be con-

GMDD
8, 5419–5435, 2015

centrated on 3 parameters of the CCSM4 parallel ocean model, namely *vconst_corr*, *vconst_2*, and *bckgrnd_vdc1*, that were earlier reported as most important by Lucas et al. Remaining parameters indicated as important in that study are either redundant or not relevant. Secondly – the machine learning methods in general, and all-relevant feature selection in particular are useful tools for analysis of influence of simulation parameters on the final outcome.

Author contributions. W. Paja perfomed most computations and drafted first version of the manuscript, M. Wrzesień and R. Niemiec performed computations and contributed to writing. W. R. Rudnicki designed experiments, and wrote the manuscript.

References

Aagaard, K., Riehle, K., Ma, J., Segata, N., Mistretta, T.-A., Coarfa, C., Raza, S., Rosenbaum, S., den Veyver, I., Milosavljevic, A., Gevers, D., Huttenhower, C., Petrosino, J., and Versalovic, J.: A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy, PLoS One, 7, e36466, doi:10.1371/journal.pone.0036466, 2012.

Ackerman, M. E., Crispin, M., Yu, X., Baruah, K., Boesch, A. W., Harvey, D. J., Dugast, A. S., Heizen, E. L., Ercan, A., Choi, I., Streeck, H., Nigrovic, P. A., Bailey-Kellogg, C., Scanlan, C., and Alter, G.: Natural variation in Fc glycosylation of HIV-specific antibodies impacts antiviral activity, J. Clin. Invest., 123, 2183–2192, 2013.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, doi:10.1023/A:1010933404324, 2001.

Buday, B., Pach, F. P., Literati-Nagy, B., Vitai, M., Vecsei, Z., and Koranyi, L.: Serum osteocalcin is associated with improved metabolic state via adiponectin in females versus testosterone in males, gender specific nature of the bone-energy homeostasis axis, Bone, 57, 98–104, doi:10.1016/j.bone.2013.07.018, 2013.

Duro, D. C., Franklin, S. E., and Dubé, M. G.: Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests, Int. J. Remote Sens., 33, 4502–4526, 2012.

Herrera, C. M. and Bazaga, P.: Epigenetic correlates of plant phenotypic plasticity: DNA methylation differs between prickly and nonprickly leaves in heterophyllous Ilex aquifolium (Aquifoliaceae) trees, Bot. J. Linn. Soc., 171, 441–452, 2013.

GMDD
8, 5419–5435, 2015

Kohavi, R. and John, G. H.: Wrappers for feature subset selection, Artif. Intell., 97, 273–324, doi:10.1016/S0004-3702(97)00043-X, 1997.

Kursa, M. B. and Rudnicki, W. R.: Feature selection with the Boruta package, J. Stat. Softw., 36, 1–13, 2010.

Kursa, M. B., Jankowski, A., and Rudnicki, W. R.: Boruta – a system for feature selection, Fundam. Inform., 101, 271–285, 2010.

Leutner, B. F., Reineking, B., Müller, J., Bachmann, M., Beierkuhnlein, C., Dech, S., and Wegmann, M.: Modelling forest α-diversity and floristic composition – on the added value of LiDAR plus hyperspectral remote sensing, Remote Sens., 4, 2818–2845, 2012.

Lucas, D. D., Klein, R., Tannahill, J., Ivanova, D., Brandon, S., Domyancic, D., and Zhang, Y.: Failure analysis of parameter-induced simulation crashes in climate models, Geosci. Model Dev., 6, 1157–1171, doi:10.5194/gmd-6-1157-2013, 2013.

Ma, J., Prince, A. L., Bader, D., Hu, M., Ganu, R., Baquero, K., Blundell, P., Alan Harris, R., Frias, A. E., Grove, K. L., and Aagaard, K. M.: High-fat maternal diet during pregnancy persistently alters the offspring microbiome in a primate model, Nat. Commun., 5, 3889, doi:10.1038/ncomms4889, 2014.

Menikarachchi, L. C., Cawley, S., Hill, D. W., Hall, L. M., Hall, L., Lai, S., Wilder, J., and Grant, D. F.: MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures, Anal. Chem., 84, 9388–9394, doi:10.1021/ac302048x, 2012.

Rudnicki, W. R., Wrzesień, M., and Paja, W.: All relevant feature selection methods and applications, in: Feature Selection for Data and Pattern Recognition, edited by: Stańczyk, U. and Lakhmi, C. J., Springer-Verlag, Berlin, Heidelberg, 11–28, 2015.

Saulnier, D. M., Riehle, K., Mistretta, T.-A., Diaz, M.-A., Mandal, D., Raza, S., Weidler, E. M., Qin, X., Coarfa, C., Milosavljevic, A., Petrosino, J. F., Highlander, S., Gibbs, R., Lynch, S. V., Shulman, R. J., and Versalovic, J.: Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome, Gastroenterology, 141, 1782–1791, doi:10.1053/j.gastro.2011.06.072, 2011.

Smith, R., Jones, P., Briegleb, B., Bryan, F., Danabasoglu, G., Dennis, J., Dukowicz, J., Eden, C., Fox-Kemper, B., Gent, P., Hecht, M., Jayne, S., Jochum, M., Large, W., Lindsay, K., Maltrud, M., Norton, N., Peacock, S., Vertenstein, M., and Yeager, S.: The Parallel Ocean Program (POP) reference manual, ocean component of the Community Climate System Model (CCSM), Tech. Rep. LAUR-10-01853, Los Alamos National Laboratory, 141 pp., 2010.

GMDD

8, 5419–5435, 2015

Strempel, S., Nendza, M., Scheringer, M., and Hungerbühler, K.: Using conditional inference trees and random forests to predict the bioaccumulation potential of organic chemicals, Environ. Toxicol. Chem., 32, 1187–1195, 2013.

UCAR: The Community Climate System Model Version 4, available at: http://www.cesm.ucar.edu/models/ccsm4.0/ (last access: 31 March 2015), 2010.

Vapnik, V. N.: The nature of statistical learning theory, Springer-Verlag New York, 2000.

GMDD

8, 5419–5435, 2015

Table 1. Summary of results. The variables indicated as important by Lucas et al. (2013) are marked with $*$, the variables that were indicated as important in the first test are highlighted in bold face. Three values are reported, the number of times the variable was deemed relevant, mean difference in AUC due to adding variable to set of variables and number of times AUC was improved by adding variable to set of variables. The first value is reported for all variables, two other are reported only for these variables that were deemed relevant significantly more often than randomised variables. The unit for Δ(AUC) is 0.0001.

Variable	**V1**$*$	**V2**$*$	V3	**V4**$*$	V5$*$	V6	Reference
# relevant	**660**	**660**	0	**44**	19	33	25 ± 9
Mean Δ(AUC)	$\mathbf{-905 \pm 80}$	$\mathbf{-749 \pm 90}$	–	20 ± 70	–	–	
# improved	**30**	**30**	–	16	–	–	
Variable	V7	V8	**V9**	V10	V11	V12	Reference
# relevant	2	17	**62**	11	3	5	25 ± 9
Mean Δ(AUC)	–	–	60 ± 70	–	–	–	
# improved	–	–	**22**	–	–	–	
Variable	**V13**$*$	**V14**$*$	V15	**V16**$*$	V17$*$	V18	Reference
# relevant	**593**	**623**	26	**67**	19	2	25 ± 9
Mean Δ(AUC)	-11 ± 60	$\mathbf{-180 \pm 80}$	–	6 ± 60	–	–	
# improved	16	**26**	–	**14**	–	–	

GMDD
8, 5419–5435, 2015

Table 2. Results of experiment 2. Average AUC obtained for all tested models, as well as examples for five interesting cases. #1 – the sample with lowest AUC from core model, #12 the sample with highest AUC obtained in the study, samples #6, #22 and #30 – samples with core model close to the mean that show variance of AUC for other models. The highest value obtained in all experiments has been highlighted in bold face.

	Sample					
Variable set	#1	#6	#22	#30	#12	Average
core	0.865	0.921	0.922	0.928	0.983	0.925 ± 0.006
core +V4	0.879	0.915	0.954	0.930	0.982	0.927 ± 0.007
core +V9	0.866	0.923	0.945	0.919	0.989	0.931 ± 0.006
core +V16	0.848	0.906	0.938	0.927	**0.990**	0.926 ± 0.007
core-V14	0.823	0.907	0.926	0.919	0.967	0.907 ± 0.007
core-V13	0.877	0.910	0.952	0.921	0.968	0.924 ± 0.006
core-V1	0.745	0.821	0.806	0.823	0.910	0.835 ± 0.007
core-V2	0.808	0.808	0.825	0.840	0.888	0.850 ± 0.009

GMDD
8, 5419–5435, 2015

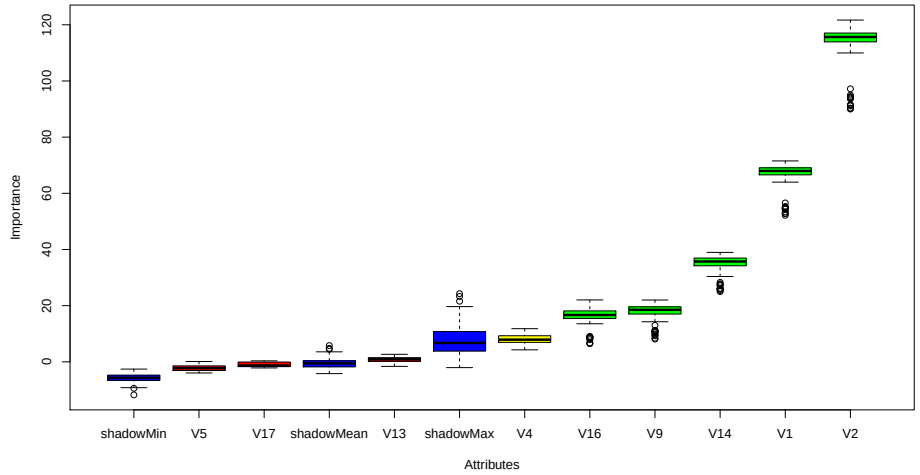

Figure 1. Summary of results of the Boruta run. Importance of the variables is shown. The variables are sorted by increasing importance. The variables coloured in green are these, which were classified as relevant. Variables coloured in red are these, which are irrelevant. The blue boxes correspond to respectively minimal (sMin), median (sMed) and maximal (sMax) importance achieved in each run by contrast variables. One can observe wide range of maximal importance values that can be achieved by random variables. In particular in many iterations it can be higher than importance of truly relevant variables.

GMDD
8, 5419–5435, 2015

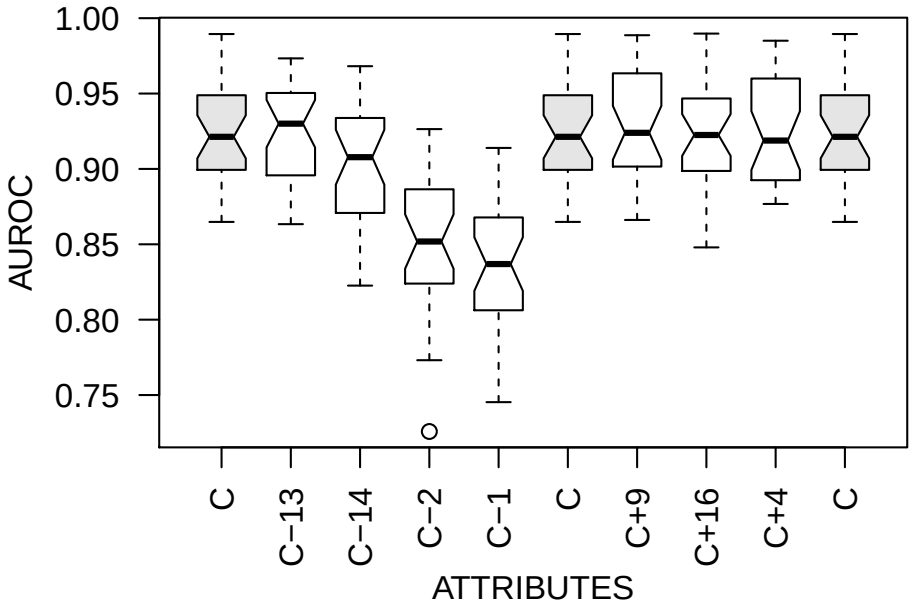

Figure 2. AUC obtained in simulations study grouped by subset of variables used for model building. The labels are coded in the following way C – core set of variables {V1, V2, V13, V14}; $C + X$ – the core set was extended by adding variable VX, where X is one of {4,9,16} ; $C - X$ – the variable VX was removed from the core set, with $X = \{1, 2, 13, 14\}$.

GMDD
8, 5419–5435, 2015

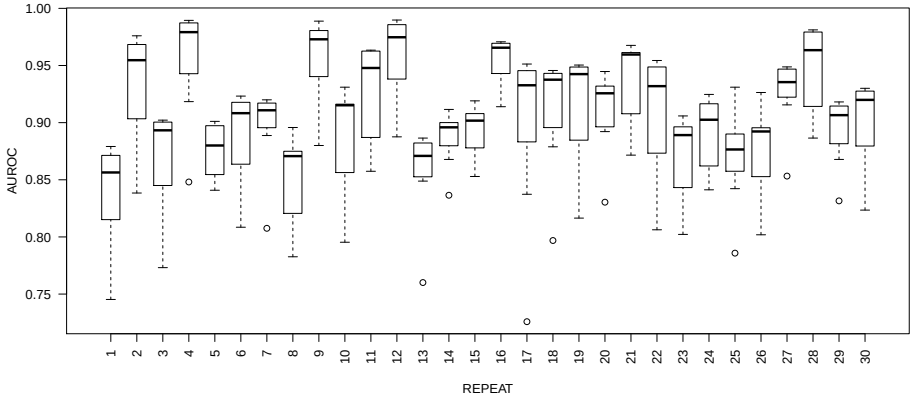

Figure 3. AUC obtained in simulations study grouped by split between training and validation set.

GMDD
8, 5419–5435, 2015