**How can the overlearning-risk be detected and handled**

**in model- and case-level based on the DIF(e)SA**

**(direct-inverse-function-symmetry-approach)?**

László Pitlik (sen) (<https://orcid.org/0000-0001-5819-0319>), Dániel Váradi (<https://orcid.org/0000-0001-9610-8566>),

László Pitlik (jun) (<https://orcid.org/0000-0002-8058-9577>), Mátyás Pitlik (<https://orcid.org/0000-0002-1991-3008>),

Marcell Pitlik (<https://orcid.org/0000-0002-8054-9364>)

e-Mails: pitlik@my-x.hu, danielvaradi140@gmail.com, ptlklszl@my-x.hu, ppk@my-x.hu, miau@my-x.hu

Kodolányi János University and MY-X research team Hungary

**Abstract**

Present status: It is well-known, that machine learning approaches can produce overlearning effects, where overlearning means a solution (a constellation of the model-parameter-set) with such kind of specific configuration which leads to a quasi-error-free approximation (concerning the learning data), but parallel, it presents a miserable estimation of further (unknown) cases. The overlearning is the result of seemingly rational objective functions. Therefore, they can not be avoided in the classic learning processes, because nobody may know in advance about the model-results in till now not-used cases. Each approach to detect overlearning effects is important for all data-analysts in order to ensure a kind of I-do-not-know-system-answer or even to explore the necessity and operationality of a finetuned estimations.

Goal/Task: The goal of the research team was to construct a new similarity-analysis-based approach for the question: How can the overlearning-risk be detected and handled in model- and case-level based on the DIF(e)SA (direct-inverse-function-symmetry-approach)? The task is trivial: We had to derive a new visualisation technique and its operationalized version to detection and as far as possible to finetuning of overlearned estimations.

Solution: The similarity analysis (or component-based object comparison: see <https://miau.my-x.hu/myx-free/index_en.php3>) is a methodology, where the problem should be described in form of an OAM (object-attribute-matrix). The similarity-based optimizing produce starting from ranking values as inputs a kind of staircase function. The most flexible version of the similarity analysis is the calculation process with doubled attribute OAMs. The doubled attribute set means a direct and an inverse (mirrored) layer of all input attributes in a parallel way. These layers are parts of the same production function at the same time (c.f. superposition, where currents in both directions can be assumed in a wire). If the sum of the direct impacts and the sum of the inverse impacts are presented in a 2D-coordinate system, there are two possibilities: either are the two curves symmetric or random-like. In symmetric cases (concerning the learning data), we can speak about massive learning results which make possible to evaluate and finetune the estimations for test cases. The evaluation of an estimation for a new test case is risk-free if the symmetry for the direct and inverse layers are given like in the learning phase. If an estimation for a test case has two not symmetric components (it means we identified an overlearning sign), then an automated correction mechanism can be started which leads to a better accuracy. If the learning phase could not be closed with a massive symmetry, then the learning problem as such could not really be learned (independent from the fitting of the learning). The lack of symmetry after a learning phase is a sign of overlearning because we do not have the possibility to identify symmetry-based risk and we do not have the chance to finetune the overlearned estimations.

Already closed experiments: The modelling based on doubled attribute sets is a context free approach. Therefore, it is possible to involve it for arbitrary problems like estimation of prices/courses, yields, number of tourists/readers, temperatures (see demo-URLs below).

Argumentations – why this technique is effective: the hidden layers (like the additive direct and inverse impacts) use the till now never observed inner symmetry of the models in order to derive diagnoses about models and/or cases. The symmetries were interpreted already in the standard hermeneutic of the similarity analyses: the lack of the case-based symmetry-effects were signs e.g., for not-gossiping (it means for I-do-not-know-system-answers) ever, where the learning phase was doubled and not the attribute set in one single learning phase. Further consistence-oriented approaches are also existing (c.f. [https://miau.my-x.hu/myx-free/index\_en.php3?\_filterText2=\*consisten](https://miau.my-x.hu/myx-free/index_en.php3?_filterText2=*consisten)).

Future: It is necessary to prepare further experiments where we can derive: how to enforce more or less symmetric results based on the variations of the inputs? Parallel, it will also be important to develop an online service: how to measure the symmetry as such in an operative way (based on the anti-discriminative version of the similarity analysis (COCO-Y0))?

Demo-URLs: e.g., <https://miau.my-x.hu/miau/293/eur_huf_holnap_integralva.xlsx>, <https://miau.my-x.hu/miau/293/stadat_turizmus_kovetkezo_negyedev.xlsx>, <https://miau.my-x.hu/miau/290/tel_2022.xlsx> (worksheets: coco7, coco7 (2), <https://miau.my-x.hu/miau/292/eur_huf_15_nap.xlsx> (worksheets: analog\_modell\_15nap, elozmenyek\_10nap), <https://miau.my-x.hu/miau/292/fagyos_napok_szama.xlsm> (worksheet: 2022(2)), <https://miau.my-x.hu/miau/294/konyvtarba_jarok_szama.xlsx> + c.f. <https://miau.my-x.hu/miau/quilt/2020/quilt2/launching2020V13/quilt_2_0_final_exam_decadence_index.docx> (figure on page#12)

# Introduction

It is well-known, that machine learning approaches can produce overlearning effects, where overlearning means a solution (a constellation of the model-parameter-set) with such kind of specific configuration which leads to a quasi-error-free approximation (concerning the learning data), but parallel, it presents a miserable estimation of further (unknown) cases. The overlearning is the result of seemingly rational objective functions. Therefore, they can not be avoided in the classic learning processes, because nobody may know in advance about the model-results in till now not-used cases. Each approach to detect overlearning effects is important for all data-analysts in order to ensure a kind of I-do-not-know-system-answer or even to explore the necessity and operationality of a finetuned estimations.

## Keywords

machine learning, overlearning/overfitting detection, consistence-oriented modelling, estimation-finetuning, symmetry-based approach, artificial intelligence, theory of aimlessness, similarity analysis, doubled input-attribute-set, ranking values

## Goals

The goal of the research team was to construct a new similarity-analysis-based approach for the question: How can the overlearning-risk be detected and handled in model- and case-level based on the DIF(e)SA (direct-inverse-function-symmetry-approach)? The task is trivial: We had to derive a new visualisation technique and its operationalized version to detection and as far as possible to finetuning of overlearned estimations.

## Tasks/answers

The similarity analysis (or component-based object comparison: see <https://miau.my-x.hu/myx-free/index_en.php3>) is a methodology, where the problem should be described in form of an OAM (object-attribute-matrix). The similarity-based optimizing produce starting from ranking values as inputs a kind of staircase function. The most flexible version of the similarity analysis is the calculation process with doubled attribute OAMs. The doubled attribute set means a direct and an inverse (mirrored) layer of all input attributes in a parallel way. These layers are parts of the same production function at the same time (c.f. superposition, where currents in both directions can be assumed in a wire). If the sum of the direct impacts and the sum of the inverse impacts are presented in a 2D-coordinate system, there are two possibilities: either are the two curves symmetric or random-like. In symmetric cases (concerning the learning data), we can speak about massive learning results which make possible to evaluate and finetune the estimations for test cases. The evaluation of an estimation for a new test case is risk-free if the symmetry for the direct and inverse layers are given like in the learning phase. If an estimation for a test case has two not symmetric components (it means we identified an overlearning sign), then an automated correction mechanism can be started which leads to a better accuracy. If the learning phase could not be closed with a massive symmetry, then the learning problem as such could not really be learned (independent from the fitting of the learning). The lack of symmetry after a learning phase is a sign of overlearning because we do not have the possibility to identify symmetry-based risk and we do not have the chance to finetune the overlearned estimations.

The tasks are trivial based on the above-descripted logic: this process should be operationalized!

## Motivations

The phenomenon of overlearning/overfitting is the most dangerous phenomenon concerning the machine learning challenges. The artificial intelligence will be better and better if the term of context-free Goodness (consistence, overfitting, etc.) can be handled step by step more effectively and efficiently. If an AI s capable of not-gossiping (it means: to derive the system-answer “I-do-not-know”, then this AI is a better AI than the nowadays still gossip-oriented alternative ones.

## Targeted groups

The war against the overlearning effects is directly relevant for each data analysts, but in an indirect way, it should be important for each customer of the analyses.

## Information-added-value

Each technique against the overfitting dangers reduces the amount/volume of general false estimations (forecasts), where not the direction of the changes could be derived, and the model do dare to gossip instead of being silent. The information-added-values accordingly to reducing estimation errors may be evaluated in a context-depending way.

# Relevant literature

The potential literature concerning modelling issues is quasi unlimited. The most relevant keywords are however: the overlearning as such, model-testing philosophies as un-matured approaches against the overlearning effects, and the similarity analysis as a specific (online, context-free) modelling tool ensuring an unlimited set of symmetry (consistence) checks for staircase-function-based solutions (c.f. [https://miau.my-x.hu/myx-free/index\_en.php3?\_filterText2=\*consisten](https://miau.my-x.hu/myx-free/index_en.php3?_filterText2=*consisten)).

The chosen reference from IBM demonstrates how monopolistic the phenomenon of the overlearning/overfitting is still interpreted:

## Overlearning/overfitting

Definition (c.f. IBM - <https://www.ibm.com/topics/overfitting>): “*What is overfitting? Overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose. Generalization of a model to new data is ultimately what allows us to use machine learning algorithms every day to make predictions and classify data.*”

Theory of the aimlessness of the modelling (c.f. <https://www.google.com/search?q=aimlessness+site%3Amiau.my-x.hu>):

Best Student Paper of Váradi, 2022: <https://miau.my-x.hu/miau2009/index_en.php3?x=e169>:

## Testing models

The trap based on IBM also nowadays - - <https://www.ibm.com/topics/overfitting>: “*How to detect overfit models: To understand the accuracy of machine learning models, it’s important to test for model fitness. K-fold cross-validation is one of the most popular techniques to assess accuracy of the model. In k-folds cross-validation, data is split into k equally sized subsets, which are also called “folds.” One of the k-folds will act as the test set, also known as the holdout set or validation set, and the remaining folds will train the model. This process repeats until each of the fold has acted as a holdout fold. After each evaluation, a score is retained and when all iterations have completed, the scores are averaged to assess the performance of the overall model.*”

Testing is a logical approach, but no test constellation can help to answer the following question: Is a particular model output here and now (in the absolute relevant present case) a result of overfitting or not?

We have also to prove whether we could estimate the model outputs without any testing? Testing as such wastes a lot of information because they could have been learned too:

Forecasting without testing: (c.f. Pitlik, 2007: [https://miau.my-x.hu/miau2009/adatlap.php3?where[azonosito]=22838&mod=l2003](https://miau.my-x.hu/miau2009/adatlap.php3?where%5bazonosito%5d=22838&mod=l2003), Barta, 2021: <https://miau.my-x.hu/miau/phd/Barta_Gergo_Theses_of_the_doctoral_dissertation_0604.pdf>):

* …

DATA-RACING-experiences (Pitlik L&M, 2022: <https://miau.my-x.hu/miau/287/DATARACING2022.docx>):

* …

## Similarity analysis

The IBM-philosophy think about the human potential against overlearning as follows- - <https://www.ibm.com/topics/overfitting>: “*How to avoid overfitting: While using a linear model helps us avoid overfitting, many real-world problems are nonlinear ones. In addition to understanding how to detect overfitting, it is important to understand how to avoid overfitting altogether. Below are a number of techniques that you can use to prevent overfitting: Early stopping, Train with more data, Data augmentation, Feature selection, Regularization, Ensemble methods*”

The recommendations from IBM do not consist of keywords like symmetry, or even consistence. The focus on testing is however one potential way but not the Way as such:

General description (Pitlik, 2014: <https://miau.my-x.hu/miau/196/My-X%20Team_A5%20fuzet_EN_jav.pdf>):

* …

# Research and findings

First, this chapter will define the DIF(e)SA technique. In this chapter, there will be listed relevant hypotheses. Finally, here and now some visualisation and interpretation scenarios will be presented:

## Definitions

DIF(e)SA: This is the name (the acronym) of the developed technique by the MY-X research team for detecting and finetuning overlearning effects. DIF(e)SA is a process where already the model building can not be realized in arbitrary ways. It needs a similarity analysis. Parallel existing direct and inverse model-input-layers on ranking levels are only given in the similarity analysis. Similarity analyses make possible to involve raw input attributes (variables) after their ranking: both based on the principle “the-more-the-more” and “the-more-the-less” in a parallel way. Regression models, where the sign of the coefficient of a variable is the results of the calculations are not capable of integrating antagonistic force fields in one single function (model). The sign of a regression coefficient determines one of the above-mentioned principle: one attribute can have only one sign, therefore one type of impact: Xi and Y will run either parallel or in an inverse way. The similarity analyses deliver however even the ratio and tendences of these antagonisms for each variable (c.f. Pitlik, 2019 - <https://miau.my-x.hu/miau/254/coco_optimum_hatasok_std_modellekkel.xlsx> – press F9 for activating randomized effects). The referenced XLS-file presents a high flexibility concerning the ceteris paribus forms between a real Xi and a real Y (see demo-URLs in the abstract). The models with a doubled attribute set let derive two parallel curves: a curve for the aggregated direct impacts and an other curve for the aggregated inverse impacts existing in the same time in the same model. In case of the objects (input-scenarios), these antagonistic layers can lead to a symmetrical or a random-like relationship. The quality of the model is depending on the level of this symmetry. If the sum of the direct impacts and the sum of the inverse impacts are presented in a 2D-coordinate system, there are two possibilities: either are the two curves symmetric or random-like. In symmetric cases (concerning the learning data), we can speak about massive learning results which make possible to evaluate and finetune the estimations for test cases. The evaluation of an estimation for a new test case is risk-free if the symmetry for the direct and inverse layers are given like in the learning phase. If an estimation for a test case has two not symmetric components (it means we identified an overlearning sign), then an automated correction mechanism can be started which leads to a better accuracy. If the learning phase could not be closed with a massive symmetry, then the learning problem as such could not really be learned (independent from the fitting of the learning). The lack of symmetry after a learning phase is a sign of overlearning because we do not have the possibility to identify symmetry-based risk and we do not have the chance to finetune the overlearned estimations.

## Experiments and hypotheses

There are following hypotheses to check:

* Is the expected symmetry on model level (after closing the learning phase) realistic at all? (yes/no)
* Is the level of this realized symmetry on model level (after closing the learning phase) measurable? (yes/no)
* Is this measuring process possible based on anti-discriminative similarity analyses? (yes/no)
* Can models be ranked based on the symmetry level in an automated way following the Knuth-principle[[1]](#footnote-1)? (yes/no)
* Is the symmetry always given for test scenarios on case-level? (yes/no)
* Is the symmetry level of one single (test)-case to measure? (yes/no)
* Can a model be evaluated as a kind of overlearned model if the symmetry level of the model (after closing the learning phase) is too low? (yes/no)
* Can a test case be evaluated as a kind of overlearned scenario if the symmetry level of the case (after closing the learning phase) is too low? (yes/no)
* Can the estimation of the output (Y) value be finetuned in case of a test scenario if the symmetry is not given? (yes/no)
* Will the estimation be more accurate (fit, consistent, realistic) after this finetuning? (yes/no)
* Is the DIF(e)SA technique context-free? (yes/partial/no – because the hermeneutics can involve context-depending layers too)

The experiments are particular models for covering quasi arbitrary analytical needs in the field of production functions: Already closed experiments: The modelling based on doubled attribute sets is a context free approach. Therefore, it is possible to involve it for arbitrary problems like estimation of prices/courses, yields, number of tourists/readers, temperatures (see demo-URLs below). Demo-URLs: e.g., <https://miau.my-x.hu/miau/293/eur_huf_holnap_integralva.xlsx>, <https://miau.my-x.hu/miau/293/stadat_turizmus_kovetkezo_negyedev.xlsx>, <https://miau.my-x.hu/miau/290/tel_2022.xlsx> (worksheets: coco7, coco7 (2), <https://miau.my-x.hu/miau/292/eur_huf_15_nap.xlsx> (worksheets: analog\_modell\_15nap, elozmenyek\_10nap), <https://miau.my-x.hu/miau/292/fagyos_napok_szama.xlsm> (worksheet: 2022(2)), <https://miau.my-x.hu/miau/294/konyvtarba_jarok_szama.xlsx> + c.f. <https://miau.my-x.hu/miau/quilt/2020/quilt2/launching2020V13/quilt_2_0_final_exam_decadence_index.docx> (figure on page#12)

Visualisation and interpretation (incl. finetuning components) of symmetry constellations:

Figure#1: … (source: own presentation)

Figure#2: … (source: own presentation)

Figure#...: … (source: own presentation)

Figure#...: … (source: own presentation)

Figure#...: … (source: own presentation)

# Results

The above-presented figures should be capable of making realistic for each human expert to accept and interpret the visual and algorithmic potential of the symmetries concerning the overfitting without any model testing actions. The hermeneutic of the symmetry relationships are a new sort of Chartism (c.f. stock exchange).

# Discussion

The symmetries for finetuning can also need context-depending interpretations: e.g., seasonal impacts (c.f. quarters of years) - parallel to the context-free interpretations (like negative impacts are excluded or maximum and minimum of the system behaviour limits are already massive exceeded, etc.). The observed and presented symmetries could only be examined in case of forecasting problems where the time as variable in a direct and/or indirect (hidden) way is always given.

Argumentations why this technique is effective: the hidden layers (like the additive direct and inverse impacts) use the till now never observed inner symmetry of the models in order to derive diagnoses about models and/or cases. The symmetries were interpreted already in the standard hermeneutic of the similarity analyses: the lack of the case-based symmetry-effects were signs e.g., for not-gossiping (it means for I-do-not-know-system-answers) ever, where the learning phase was doubled and not the attribute set in one single learning phase. Further consistence-oriented approaches are also existing (c.f. [https://miau.my-x.hu/myx-free/index\_en.php3?\_filterText2=\*consisten](https://miau.my-x.hu/myx-free/index_en.php3?_filterText2=*consisten)).

# Conclusions

It is possible to evaluate models (estimations for one or more given input-constellation(s)) without testing. It is possible to create models where the overfitting effects can be visualized and finetuned. It is not possible to make the problem universes deterministic. There can be parallel solutions in a particular decision situation.

# Future

It is necessary to prepare further experiments where we can derive: how to enforce more or less symmetric results based on the variations of the inputs? Parallel, it will also be important to develop an online service: how to measure the symmetry as such in an operative way (based on the anti-discriminative version of the similarity analysis (COCO-Y0))?

# References

1. [https://miau.my-x.hu/miau2009/index\_tki.php3?\_filterText0=\*knuth](https://miau.my-x.hu/miau2009/index_tki.php3?_filterText0=*knuth) [↑](#footnote-ref-1)