Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220406

Supporting AI-Explainability by Analyzing Feature Subsets in a Machine Learning Model

Lucas PLAGWITZ[a,1] Alexander BRENNER[a], Michael FUJARSKI[a], and Julian VARGHESE[a]
[a] *Institute of Medical Informatics, University of Münster, Germany*

Abstract. Machine learning algorithms become increasingly prevalent in the field of medicine, as they offer the ability to recognize patterns in complex medical data. Especially in this sensitive area, the active usage of a mostly black box is a controversial topic. We aim to highlight how an aggregated and systematic feature analysis of such models can be beneficial in the medical context. For this reason, we introduce a grouped version of the permutation importance analysis for evaluating the influence of entire feature subsets in a machine learning model. In this way, expert-defined subgroups can be evaluated in the decision-making process. Based on these results, new hypotheses can be formulated and examined.

Keywords. explainable AI, permutation importance, grouped variable analysis

1. Introduction

Whether in clinical research or risk factor calculations, machine learning algorithms can be found in many medical fields. This results in important challenges for the usage of these algorithms - especially regarding the ethical background - like regulatory aspects, interpretability, or interoperability [1]. For this work, we present a particular perspective on interpretability of feature contributions. Namely, this should not be seen as just a means of application, but rather a way to better understand the underlying problem. Our focus is on the combination of humanly understandable feature subgroups and their respective importance scores.

Feature importance analysis is a major component in the examination and interpretation of machine learning models. The question arises on which basis a problem was solved, e.g., in classification, why the decision for a class was made the way it was. The methodology behind such an analysis can vary widely depending on the problem and model. We focus in the following on the permutation importance analysis [2]. Here, the variation of predictive accuracy is analyzed based on the test set. The feature under investigation is permuted and the prediction is compared with the unmodified one. Therefore, this method is not dependent on a specific implementation of the prediction process and thus can be applied to any feature-based supervised machine learning model.

[1] Corresponding Author, Lucas Plagwitz, Institute of Medical Informatics, University of Münster, Münster, Germany; E-mail: lucas.plagwitz@uni-muenster.de.

This allows a wide range of application of this method, even for more complicated architectures such as deep neural networks.

In this paper we are concerned with a particular adaptation of the permutation importance analysis. Many data sets consist of features that can be grouped to clinically meaningful subsets. An arrangement into such subsets does not have to be unambiguous at all but still relevant for the domain. We present a methodology to determine the information gain for each of these groups. In particular, for tree-based classifications, analyses on grouped feature importance and terms like group permutation importance measures have been described previously [3]. Their research supports the theoretical aspect of our application. As a complement to this, we intend to focus more on interpretation - by using expert-defined feature subsets - than optimization of the model. To this end, we present an explicit algorithm that uses cross-validation to provide a fast, stable, and meaningful analysis based on the test performance. In doing so, we shed some light on a black box prediction. In addition to the method itself, we focus on an application in neuroimaging where we evaluate the impact of a brain region depending on the target variable.

2. Methods

In the context of grouped characteristics, it is not advisable to simply add up individual feature importance scores. The reasons depend on the method of calculating scores. For example, in the univariate permutation importance analysis, it is assumed that all other features are known at the time of permutation. This does not provide adequate insight into the influence of a specific feature group. Other feature importance methods, such as a tree-based calculation, tend to overfit, often boosting features that contribute hardly any information to the problem [4]. Thus, a sum of these feature weights overestimates the group with substantially more features, while the mean or median discriminates against this group.

For this reason, we adjust the calculation as follows: First, a machine learning model is trained based on the training set and a feature subset is specified. Then, a predefined number $T \in \mathbb{N}$ of test set permutations are created. Here, it is important that only the features within the subset are permuted. Predictions are then created based on these modified test sets. By matching the correct labels, the change in performance is estimated (averaged over all permutations). This process is illustrated in Figure 1 and is repeated for every feature subset, starting with the existing trained model.

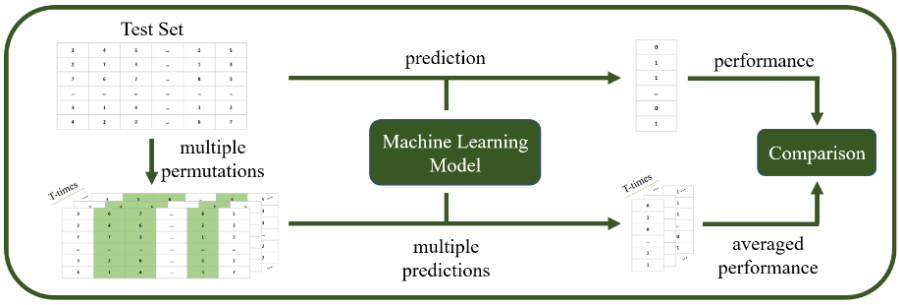

Figure 1. The permutation process of a specific feature subset is shown. The columns that are affected by the permutation are marked in green, the remaining columns remain unchanged.

Algorithm 1 Grouped Permutation Importance

Input: untrained model M; data X; targets y; k different feature subsets g_i; repeated permutation count T; cross-validation object CV, scoring-metric $m(\cdot,\cdot)$
Output: grouped permutation importance GPI $\in \mathbb{R}^k$

GPI $= 0^{(k)}$
for $(idx_{train}, idx_{test}) \in \text{CV}.split(X)$ **do**
  $\text{M}.fit(X[idx_{train}], y[idx_{train}])$
  $y_0 = \text{M}.predict(X[idx_{test}])$
  for $i = 1,\ldots,k$ **do**
    for $t = 1,\ldots,T$ **do**
      create permutation matrix P_t
      Set $y_{g_i}^{(t)}$ to the prediction from the column-wise permutation P_t in columns g_i of $X[idx_{test}]$.
      $\text{GPI}[i] = \text{GPI}[i] + (m(y[idx_{test}], y_0) - m(y[idx_{test}], y_{g_i}^{(t)}))/(|cv\_splits| \cdot T)$
    end for
  end for
end for

To stabilize the validity, this process is repeated over several cross-validation folds. The final value is then composed of the mean between these folds. Algorithm 1 summarizes the whole method. The corresponding code is published as open source[2]. The grouped permutation importance (GPI) describes the direct influence compared to the respective overall performance. Thus, it represents the absolute information gain per group. However, these analyses are particularly interesting in comparison with the same data set but different target variables. Since the maximum performance of a model depends strongly on the underlying target, a relative value should be preferred to an absolute one in these cases. For this purpose, the GPI is set in relation to the unpermuted performance to be able to analyze a relative information gain.

3. Experiment: Brain region impact depending on the target variable

The explanation of machine learning algorithms in the field of neuroimaging is a widespread analysis. One is interested in specific brain markers to gain a more detailed insight into the functionality of the brain. In the following, we see that depending on the target variable, expert-defined brain areas are incorporated as feature groups into the machine learning model on very different weights.

Based on a predefined atlas, the measured voxels were clustered and averaged into feature subgroups that we can evaluate by applying Algorithm 1. For the simulation, we use the public data set OASIS [5]. In addition to the T1-weighted magnetic resonance imaging (MRI) scans of 403 subjects, this data set provides three target variables: age, biological sex, and the clinical dementia rating (CDR), which is used for the diagnosis

[2] The code is available on GitHub and the installation is possible via pip: pip install git+https://github.com/lucasplagwitz/grouped_permutation_importance

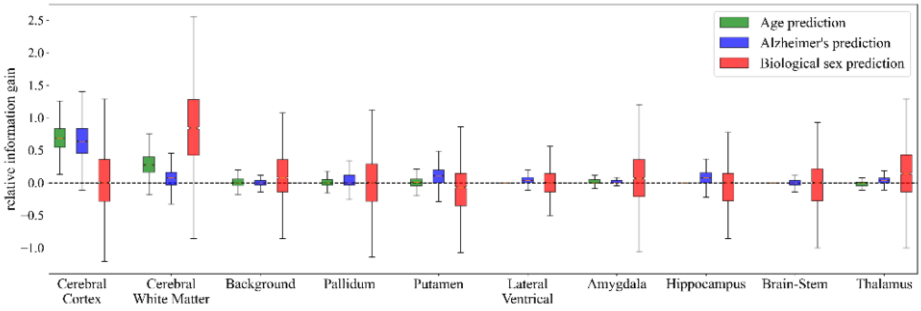

Figure 2. Brain region influence depending on the target variable. The displayed relative scores are calculated based on the balanced accuracy score. The following performances depending on the classification task were obtained: age 87%, Alzheimer's 82%, and biological sex prediction 72 %.

of Alzheimer's disease. For simplicity, we change the target variables to binary values. We only distinguish between CDR zero or greater than zero for Alzheimer's prediction and age more or less than 50 for an age-based classification task. By applying the Harvard-Oxford 2mm Subcortical Atlas, we obtain several volumes for different brain areas as seen in Table 1.

Cerebral Cortex	Cerebral White Matter	Putamen	Pallidum	Background	Lateral Ventrical
4	13	2	2	2	1

Brain-Stem	Amygdala	Hippocampus	Thalamus
1	2	2	2

Table 1. Havard-Oxford (Subcortical) Atlas – volumes per region.

All atlas calculations are based on the Python package nilearn [6]. We then determine the GPI of each brain region given a balanced support vector classifier, 5-fold cross-validation (randomized in 10 replicates), and a permutation value T = 100 by using the Python package scikit-learn [7], which is visualized in Figure 2. It is noticeable that the information about age and Alzheimer's disease is mainly located in the cerebral cortex. However, the cerebral white matter contains much more information about age. In contrast, the prediction of biological sex, this is based mainly in the white matter and hardly in the cerebral cortex. Moreover, the influence of the putamen as well as the hippocampus should be emphasized. While these regions provide no additional benefit for the prediction of age or biological sex, a positive impact on predicting Alzheimer's disease can be determined. Findings such as these could lead to a better differentiation of Alzheimer's disease from the normal aging process.

4. Discussion

We have seen that an analysis of predefined feature subsets can provide new insights into brain feature discrimination. This expands the utility of a classification algorithm from a final output to an investigative process. The consideration of other data structures is equally conceivable: As an example, time series usually offer the possibility to divide them into subsets. On the one hand, there are multivariate time series, which represent several signals over time. Medical examples are the electrocardiogram (ECG) or

electroencephalography (EEG), where information is obtained from various leads or multiple electrodes on the scalp. If we build a machine learning model that extracts features from each signal independently (e.g., sliding-window approaches), we are able to describe the information gain of every signal. With this knowledge, the cardiologist's focus could be directed to a specific lead in the ECG - or the neurologist to a specific EEG-channel. On the other hand, a subdivision into phases is also conceivable for univariate time series.

However, it is precisely at this point that a limitation of our method becomes apparent. The features must be extracted independently of the subsequences. Modern end-to-end algorithms, such as convolutional neural networks, cannot be measured with the presented approach. Furthermore, a more in-depth analysis of the method based on a comparison to alternative algorithm (e.g., grouped SHAP values), different metrics, or multiclass problems are still pending for the future.

Nevertheless, the additional benefit of analyzing the importance of entire subsets through feature-based methods should not be ignored. In contrast to time series, all groupable types of data are conceivable, whether a questionnaire according to categories or a clinical examination according to organ parameters.

5. Conclusion

We presented the grouped permutation importance to achieve a better understanding of the underlying problem by examining subsets of features. In this context, better localization of information in the data produces new insights. Through an example from brain research, we have shown how different brain regions are involved in a decision-making process depending on the target variable. In addition, many applications are conceivable since, especially in medicine, data sets usually consist of a large number of characteristics that can be grouped together in clinically meaningful categories. Determining their impact on a predictive algorithm opens entirely new possibilities in understanding, detecting, and treating diseases.

References

[1] Varghese J. Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. Visceral medicine, 36, 443-449, 2020.
[2] Breiman L. Random Forests. Machine Learning, 45(1), 5-32, 2001.
[3] Gregorutti B, Michel B, Saint-Pierre P. Grouped variable importance with random forests and application to multiple functional data analysis. Computational Statistics and Data Analysis, 90, 15-35, 2015.
[4] Loughrey J, Cunningham P. Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets. Research and Development in Intelligent Systems XXI, 33-43, 2005.
[5] Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience, 19, 1498-1507, 2007.
[6] Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics, 8:14, 2014.
[7] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830, 2011.