APPLICATION OF ENSEMBLE MACHINE LEARNING FOR CLASSIFICATION PROBLEMS ON VERY SMALL DATASETS

Ognjen Pavić[1], Lazar Dašić[1], Tijana Geroski[2,3], Marijana Stanojević Pirković[4], Nenad Filipović[2,3]

[1] Institute for Information Technologies Kragujevac, University of Kragujevac
e-mail: opavic@kg.ac.rs , lazar.dasic@kg.ac.rs
[2] Faculty of Engineering, The University of Kragujevac, Sestre Janjica 6, 34000 Kragujevac
e-mail: tijanas@kg.ac.rs , fica@kg.ac.rs
[3] Bioengineering Research and Development Center (BioIRC), Kragujevac
[4] Faculty of Medical Sciences, University of Kragujevac
e-mail: marijanas14@gmail.com

Abstract:

Machine learning is one of the most widely used branches of artificial intelligence in recent years. It is most commonly used for solving classification or regression problems through the utilization of supervised learning approaches. Machine learning models require high quality and a sufficient quantity of data to produce good results. This paper investigates an approach which incorporates ensemble learning through the aggregation of multiple machine learning models for the purposes of increasing prediction capabilities in cases in which a very limited amount of data is available for training. The ensemble model was trained on a patient fractional flow reserve biomarker dataset and had a goal of classifying patients into risk classes based on their risk of suffering an acute myocardial infarction. The ensemble model was comprised of multiple random forest classification models which were trained with different combinations of training and test data to improve the prediction accuracy over the use of a single random forest model. Final ensemble achieved a prediction accuracy of 71.3% which was an immense improvement over the 36% prediction accuracy of a single random forest classification model.

Key words: Machine learning; Classification; Risk assessment; Random forest; Ensemble

1. Introduction

Machine learning is a branch of artificial intelligence which utilizes a variety of learning algorithms for the creation of mathematical or logic-based models, with a goal of representing connections and dependencies within a dataset, so that they can be utilized in the future on new data points. The most common use of machine learning is solving classification and regression problems through supervised learning approaches. Classification and regression machine learning algorithms are often used for but are not limited to solving problems including the need for semantic classification, risk stratification, uncovering hidden knowledge, uncovering trends over time, future prediction of values through data extrapolation etc., and along with deep learning approaches are most commonly utilized in the fields of medicine, economics, computer vision and robotics.

Machine learning is a powerful tool that can be used for automation and streamlining of complex tasks and for the uncovering of hidden knowledge which can be utilized for further research in the future. However, machine learning requires data that is descriptive enough for a model to be created while also being large enough to minimize errors efficiently and avoid model underfiting. A machine learning model trained with bad data cannot produce satisfying results; on the other hand, an abundance of data is not always available for model training. In this paper we have created an ensemble of machine learning models whose purpose was to increase the predictive capabilities in contrast to using a single model with a very small amount of available data.

Ensemble learning implies the creation of multiple less complex machine learning models which work together to come to a conclusion [1]. Ensemble machine learning models can be created by aggregating multiple models of lesser complexity created by the same algorithm, but can also consist of models created using several different algorithms [2].

2. Materials and methods

2.1. Dataset

The dataset consisted of clinical data gathered from patients in the form of biomarkers, as well as descriptive data which contained information on the primary and follow up diagnosis and definitions of location and degrees of stenosis and lesions in the right coronary artery, left anterior descending artery and the left circumflex artery. The dataset contained information about 112 patients, however only 17 of these patients had known ground truth values regarding the risk classification target. The main goal was to create a machine learning model that is capable of classifying patients into appropriate risk classes based on their risk of suffering acute myocardial infarction, utilizing the patients for whom the class was known and apply the classification on the rest of the available data.

The available data had multiple problems which needed to be addressed during data preprocessing. The set contained instances of missing values which were filled in using conventional approaches depending on the type of missing data. Namely, numeric data was filled in using the mean value of the appropriate column and binary data was filled in using the most common binary value contained within that column. The second problem which existed within the data was the descriptive nature of certain features. Descriptive data containing information on stenosis and lesion in coronary arteries needed to be translated into a numeric value. This translation was conducted as following:

- Data that contains percentile values for the narrowing of the observed artery was translated as a numeric sample corresponding to the percentage value.
- Data that contains an approximation of the narrowing in the form of a range of values was translated as a numeric sample corresponding to the average value of the observed range.
- Data that does not contain percentage values of the narrowing, but has the indication that the narrowing is not substantial was translated as if it held information about a narrowing of 10%
- Data that does not contain percentage values of the narrowing, but has the indication that the narrowing is very minor was translated as if it held information about a narrowing of 5%
- Data that does not contain percentage values of the narrowing but indicates an orderly arterial lumen was translated as if there was no narrowing at all.
- Data that does not contain any indication of the size of the narrowing, nor does it contain the previously mentioned phrases with which the narrowing was estimated was not translated at all but was approximated as a mean value of all of the other translated values.

2.2. Individual model training

Individual machine learning models were constructed by using the random forest classification algorithm. Because 17 labeled data samples were not enough to build a comprehensive test set, multiple random forest models were trained using different combinations of training and test data.

Random forest classification algorithm produces classification models which are ensemble models on their own, by training multiple decision tree models using randomly selected subsamples of training and test data for each one. Each of the created decision tree model comes to its own conclusion and the random forest model chooses the output value which is present in the highest number of cases. Each random forest model was created using 50 decision trees without constraint in regard to minimum samples required for creating branching nodes and leaves.

2.3. Ensemble model creation

The available dataset was split into every possible configuration of training and test data with a 16:1, 15:2, 14:3 and 13:4 train-test split. For every existing combination of training and test data, a single random forest classification model was trained. Although many classification models had to be trained, it was not efficient for every one of them to be a part of the final ensemble, because the decision-making process would take a very long time. Only certain models, which performed the best on their own combinations of data, were selected to be the part of the final ensemble. The selection criteria were set based on their achieved prediction accuracy and the number of data samples used for testing. Models which were tested using a single data sample were selected if they achieved 100% prediction accuracy. Models evaluated using 2, 3 or 4 test samples were selected if they achieved a minimum of 50%, 66% and 75% prediction accuracy respectively.

The final ensemble model was then constructed through the sequential use of the selected models. Each of the selected models would generate an output class and the entire ensemble then chooses the class which was chosen the highest number of times.

2.4. Results

The final ensemble model was evaluated using the prediction accuracy metric. It achieved a prediction accuracy value of 71.32% which was much higher than the accuracy achieved by using a single random forest model which was at 36% after fine tuning.

Average performances for models trained with different configuration of training and test sets are shown in Table 1.

Table 1. Classification accuracy metrics

train:test split	Mean prediction accuracy
16:1	100%
15:2	56%
14:3	69%
13:4	76%
Final model	71.32%

3. Conclusions

Creating machine learning models to solve classification or regression type problems will yield less than satisfying results if the dataset used for their training does not contain high quality and a high quantity of data.

In this paper we were met with the challenge of a very small dataset available for model creaion. Although this challenge is impossible to overcome without using a better suited dataset, we were able to majorly improve the final results. Results were improved throguh the creation of a large number of less complex machine learning models and their agregation into an ensemble model.

There exist multiple other approaches to creating an ensemble model mainly through the use of multiple different machine learning algorithms during the creation of lesser complexity models, and through the change in the inner workings of the final decision selection process, but the described approach achieved a major increase in overall model performance.

In conclusion we managed to achieve satisfying results through the use of ensemble learning and improve the quality of model outputs when compared to the use of a less complex model. The final results cannot be further improved through model tuning without the expansion of the available dataset through the introduction of more labeled data.

Acknowledgements:

The research was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, contract number [451-03-47/2023-01/200378 (Institute for Information Technologies Kragujevac, University of Kragujevac)]. This research is also supported by the project that has received funding from the European Union’s Horizon 2020 research and innovation programmes under grant agreement No 952603 (SGABU project). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains.

References

[1] R. Polkar, "Ensemble learning," *Ensemble Machine Learning, Springer.*, vol. 1, no. 1, pp. 1-34, January 2012.

[2] Omer Sagi and Lior Rokach, "Ensemble learning: A survey," *WIREs data mining and knowledge discovery*, vol. 8, no. 4, February 2018.