Logical analysis of data using linear approximation and heuristic algorithms for gene expression-based diagnostics

Maria Bartosh[1], and *Igor* Masich[1,2*]

[1]Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy av., Krasnoyarsk, 660037, Russian Federation
[2]Siberian Federal University, 79, Svobodny av., Krasnoyarsk, 660041, Russian Federation

Abstract. This research aims to develop a methodology that combines logical analysis of data with a white box model to predict the progression of chronic diseases. Such diseases represent a serious health problem, and accurate prediction and management are essential to improve patients' quality of life. Current machine learning methods such as deep learning often have high accuracy, but their solutions are 'black boxes', making them difficult to understand. The research combines the best aspects of both methods to create more accurate and interpretable models for predicting the progression of chronic diseases. The methodology developed is expected to contribute to informative decision-making in medical practice, enrich knowledge in medical research and improve the quality of care for patients with chronic diseases.

1 Introduction

A white box model is a model in machine learning or statistics that is characterized by a high degree of interpretability and explainability. This means that users can easily understand how the model makes decisions and what factors influence the results.

The white box model remains relevant and important, especially in medical practice and research [1]. One of the main reasons is the interpretability of the results. The model provides clear rules and patterns, which is important for clinicians, researchers and patients. In medical practice, it is important to understand why a certain diagnosis and treatment decision was made. With this approach, physicians and patients can feel more confidence in decisions made based on interpretable models. White box models can justify decisions based on specific attributes and data, which is a great advantage in scientific research. In medical research and clinical trials, such models can be used to determine the effectiveness of new diagnostic and treatment methods. Interpretable results allow for a more accurate assessment of the impact of new approaches. Despite the current development of "black boxes" (e.g., deep neural networks), white box models will always have a place in medical practice and research due to their interpretability and applicability in cases where transparency and understanding of decision-making is important.

* Corresponding author: i-masich@yandex.ru

Interpretable machine learning models can help in the management of diseases such as diabetes [1], Alzheimer's disease, heart disease, autoimmune diseases, oncology and many others. It is important for patients with these conditions to make informed and educated decisions about how to manage the disease and maintain quality of life. The approach to treatment can be highly individualised. White box models can take into account the unique characteristics of each patient and suggest the most appropriate treatment and disease containment strategies. Interpretable models can also be used to predict disease course and monitor disease progression, allowing clinicians to decide whether treatment adjustments are necessary.

2 Problem statement

Chronic diseases represent a serious health problem, and their timely prediction plays a key role in improving the quality of life of patients. In this context, it is important to develop a new method for predicting the development of chronic diseases that combines a method of logical data analysis with a white box model. Although various prediction methods exist, there remains a need for more accurate and interpretable models that allow understanding and explanation of physicians' decisions. White box models provide interpretability but have limitations in prediction accuracy. Logical analysis of data can help improve the accuracy and explainability of models, which is particularly important in medical practice.

3 Research questions

This research used the GSE42568 dataset containing gene expression profiles of patients diagnosed with breast cancer as the basis [2].

Table 1. GSE42568 dataset.

type	1007_s_at	1053_at	117_at	121_at
normal	7.9442	5.2569	4.9346	6.6084
normal	8.8840	5.3316	4.9048	7.2040
tumoral	10.0278	5.8531	4.9042	6.5938
tumoral	9.2955	5.5814	4.9910	6.5508

A sample of 101 breast biopsy specimens from patients diagnosed with breast cancer was analysed. The age of the patients at the time of diagnosis ranged from 31 to 89 years, with a mean age of 58 years. Of these, 26 were younger than 50 years of age, and 75 women were 50 years of age or older. The size of the tumours ranged from 0.6 cm to 8.0 cm, with a mean size of 2.8 cm. In 18 cases, the tumour was T1 (less than 2 cm in maximum dimension), 80 cases were T2 (2-5 cm), and 3 cases were T3 (more than 5 cm). 11 specimens were grade 1 tumour differentiation, 37 were grade 2, and 53 were grade 3. 65 specimens were estrogen receptor (ER) positive and 33 were ER negative. Information on ER status was not available for 3 patients. No metastasis to axillary lymph nodes was detected in 43 patients, and 58 patients had metastasis. The maximum follow-up period was 3026 days and the mean follow-up period was 1881 days. There are also 15 normal tissue samples in the set. The sample contains 116 observations and 54677 features. The dataset includes 54676 real features and 1 dependent attribute describing the observation class. The sample is free of omissions and interferences.

4 Purpose of the study

Patients and clinicians need more informative and interpretable predictive decisions about chronic disease progression in order to choose the most appropriate treatment and disease control pathway. Modern machine learning techniques such as deep learning often have good predictive accuracy, but their decisions are "black boxes" and difficult to interpret. On the other hand, logical analysis of data allows for the construction of more understandable rules and patterns, but has its limitations.

The aim of this research is to develop a methodology that combines logical analysis of data and the white box model to improve the accuracy and interpretability of chronic disease prediction. The development of such an approach to data analysis may have broad potential in research and clinical medicine, enriching the understanding of factors influencing disease progression and improving treatment decisions.

5 Research methods

Interpretable machine learning is an approach that provides insight into how an algorithm makes decisions. This approach takes into account the interpretability of models during their creation, in contrast to the traditional methodology, where models can be very complex and incomprehensible to humans.

Interpretable machine learning uses techniques that help explain how a model makes decisions, how reliable those decisions are, and what factors have the greatest influence on a decision. This can help improve understanding and data-driven decision making, especially when models are used in critical areas such as medicine or finance.

There are several popular interpretive machine learning algorithms, one of them logistic regression is used for binary classification and can be interpreted using the coefficients of the logistic function. This allows us to determine how each predictor affects the probability of belonging to a particular class. Decision trees are built from a series of decisions and their structure can be interpreted graphically. A random forest is an ensemble of decision trees. By aggregating the decisions of an ensemble of trees, the interpretability can be improved and the probability of overfitting can be reduced. Gradient boosting, which is used for classification and regression tasks, can also be used to build a model. It provides methods for estimating the importance of features and facilitates interpretation. Gradient boosting with decision trees such as CatBoost and LGBM can achieve high accuracy while maintaining interpretability.

Logical analysis of data (LAD) [3] is not itself a white box model, but it is a method of analysing data that pays particular attention to the interpretation of results. This approach involves the use of formal logic and rules to extract knowledge and patterns from data. It can be used in a variety of fields, including medicine, to analyse and interpret medical data. A central concept in LAD is patterns, or rules, which play an important role in classification, ranked regression, clustering, subclass detection, feature selection and other problems. The research area was defined and initiated by Peter L. Hammer who has been a catalyst for directed research for decades, his vision and efforts have helped the methodology move from theory to data analysis applications, reach maturity and be successful in many areas.

The process of working of the algorithm of logical analysis of data starts with preliminary data preparation. The sample may contain a large number of attributes, which can complicate further work of the method in terms of the model training process. On this basis, it is necessary to reduce the number of attributes, leaving those that have a greater influence on the separation of observations into classes [4]

Binarization of attributes is a key step in logical analysis of data. In this research, the procedure for obtaining binary features is the conversion of real attribute values into binary

values depending on the frequency of class change. Therefore, one real attribute can correspond to 1 to 30 binary attributes.

The search for a reference set of features allows us to identify the attributes that contain the greatest number of differences between observations of different classes. To implement this process, a greedy algorithm is used to search for a reference set of attributes to find the minimum set of variables that can separate observations based on the target attribute.

An optimisation problem is used to generate logical patterns in the data. This approach finds attribute dependencies to classify observations. To begin with, it is necessary to generate an optimisation problem, i.e.:

1. define the criteria/variables of the model,
2. define the objective function for the model,
3. define the constraints of the model.

In the case of two classes K+ and K-, for any observation $\alpha \in K^+$ and variables y_j equals to 1 if j attribute is included to the pattern (and 0 otherwise), the optimization problem is [5]

$$\sum_{\beta_j \in K^+} \prod_{j=1,\beta_j \neq \alpha_j}^{n} (1 - y_j) \to \max_{Y} \quad (1)$$

$$\sum_{j=1,\gamma_j \neq \alpha_j} y_j \geq 1, \forall \gamma \in K^- \quad (2)$$

There is a method of approximate calculation used to construct a linear model that best fits a set of data - the best linear approximation (BLA) [5]. This method is an optimisation problem, namely minimising the objective function described in formula 3. To start the calculation, we need to find the number of differences between observations of positive class and negative class, as shown in formula 4.

$$L(\prod_{i \subseteq S} u_i) = \sum_{j=1}^{n} \left(\sum_{\beta \in K^+, \beta_j \neq \alpha_j} \frac{1}{2^{\omega(\beta)-1}} \right) y_j \to \min_{Y} \quad (3)$$

$$\omega(\beta) = |\{j: \beta_j \neq \alpha_j\}| \quad (4)$$

One of the advantages of using BLA is its simplicity, comprehensibility and efficiency when dealing with large amounts of data. With BLA, the problem of finding maximal patterns is reduced to an optimization problem of integer linear programming, which is solved using linear solvers.

The generated patterns are parts of the decision function, when combined we obtain a classifier.

6 Results and discussion

The logical analysis of data algorithm is implemented in the Python programming language. The current dataset contains a large number of significant features, namely 54676. Analysing such a dataset is a long and resource-intensive process, so it is necessary to reduce the number of significant features for classification. The determination is done by applying the envelope eccentricity metric. By default, after pruning, the number of significant features is 50.

The step of finding logical regularities or patterns is necessary to find hidden relationships between features in order to build a more accurate classifier. This step is implemented using the best linear approximation method. This algorithm is implemented using the SciPy library in Python [6].

The final step of the LAD algorithm is to build a classifier based on the generated patterns. The decision function can be constructed using a decision tree, but in the current implementation, the greedy algorithm is applicable. The method chosen is to sequentially remove the covered observations and the pattern that covers them from the coverage table. As the target function, we define the function of maximising the coverage of the observations of the training sample. Therefore, the logical analysis of data algorithm constructs a decision function consisting of 6 rules: 244057_s_at < 5.625 or 236642_at ≥ 5.428 or 221104_s_at < 3.474 for class "tumoral" and 244057_s_at ≥ 5.625 or 236642_at < 5.428 or 221104_s_at ≥ 3.474 for class "normal".

The selected dataset is unbalanced, as the "normal" class contains 15 observations and the "tumoral" class contains 101 observations. Therefore, the accuracy metric is not valid. Precision and recall are class-independent, so they can be applied to unbalanced datasets. These metrics evaluate each class separately. However, these scores can be combined into one. The F-measure combines the precision and recall metrics.

During the comparison, we used algorithms for building classifiers that have implementations for solving machine learning problems: Decision Tree [7], Naive Bayesian Classifier [8], Random Forest [9], kNN [10], Gradient Boosting [10], MLP based on neural networks [11], SVM [10], LGBM [12].

Table 2. Comparison of the effectiveness of machine learning algorithms.

Algorithm	Precision	Recall	F-score
LAD	100%	100%	100%
Gradient Boosting Classifier	98.14%	100%	99.05%
Random Forest Classifier	97.27%	100%	99.53%
Naive Bayesian Classifier	100%	100%	100%
kNN	99.09%	100%	99.53%
MLP Classifier	87.07%	100%	93.09%
SVM	99.09%	100%	99.53%
Decision Tree Classifier	97.27%	98.1%	97.67%
LGBM Classifier	96.84%	93.24%	94.97%

The results show that logical analysis of data is not inferior in accuracy, and sometimes even surpasses other known machine learning algorithms.

7 Conclusion

A methodology for logical analysis of data was investigated and implemented using a greedy heuristic algorithm and linear approximation of the objective function together with an algorithm for solving an integer linear programming problem. A classifier was built to predict the development of cancer based on gene expression data. At the moment, the constructed set of rules successfully classifies observations belonging to one of two types of breast tissue, namely healthy tissue and tumor. This algorithm can be used in medical institutions, for example, in cancer centers, as a decision support system. With the help of this technology, oncologists will be able to prescribe the optimal course of treatment or adjust an existing one, taking into account the patient’s prognosis for the development of the disease. The resulting classifiers are compact, have high accuracy and good interpretability.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121).

References

1. G. Cappon, F. Prendin, A. Facchinetti, G. Sparacino, S. D. Favero, IEEE Transaction on Biomedical Engineering **70**, 3105-3115 (2023)
2. C. Clarke, S.F. Madden, P. Doolan, H. Joyce, S. Aherne, L. O' Driscoll, W.M. Gallagher, B. Hennesy, M. Moriarty, J. Crown, S. Kennedy, M. Clynes, Carcinogenesis **34**, 2300-2308 (2013)
3. G. Alexe, S. Alexe, T.O. Bonates, A. Kogan, Annals of Mathematics and Artificial Intelligence **49**, 265-312 (2007)
4. G. Alexe, S. Alexe, P.L. Hammer, B. Vizvari, Annals of Operations Research **148**, 189-201 (2006)
5. T.O. Bonates, P.L. Hammer, A. Kogan, Discrete Applied Mathematics **156**, 845-861 (2008)
6. Q. Huangfu, J.A.J. Hall, Mathematical Programming Computation **10**, 119-142 (2018)
7. S. Zhifang, L. Yi, Journal of Physics: Conference Series **1693**, 012219 (2020)
8. G. Ou, Y. He, P. Fournier-Viger, J.Z. Huang, Appl. Sci. **12**, 10443 (2022)
9. C. Lauw, H. Hairani, I. Saifuddin, J. Guterres, M. Huda, M. Mayadi, International Journal of Engineering and Computer Science Applications (IJECSA) **2**, 59-64 (2023)
10. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Journal of Machine Learning Research **12**, 282-2830 (2011)
11. S. Samarpita, R. Satpathy, P. Kumar Mishra, A. Narayan Panda, EAI Endorsed Trans Perv Health Tech **9**, (2023)
12. Y. Hong, X. Zhang, BMC Public Health **22**, 2027 (2022)