UDC 004+616.43

ENSEMBLE METHODS FOR SOLVING PROBLEMS OF MEDICAL DIAGNOSIS

Anastasia A. Grigoreva[1],
aag@isti.edu; isaevaaa@ex.istu.edu

Andrei I. Trufanov[1],
ait@isti.edu; troufan@istu.edu

Stanislav V. Grigorev[1],
svg@isti.edu

[1] Irkutsk National Research Technical University,
83, Lermontov street, Irkutsk, 664074, Russia.

The authors proposed the consolidating method for analyzing series of observations based on a fitted model of a mixture of catalysts of the main components, which makes it possible to study any number of markers. Contrasting the longitudinal approach, it eliminates the need to connect regression analysis methods with their own uncertainties when choosing particular models. The consolidating method allows obtaining an original result in the subject area of early diagnosis of a disease: all options for using markers demonstrate an increase in classification accuracy with an increase in the length of a series of examinations.

Key words: *ensemble learning, data aggregation, statistical classification, model accuracy, medical diagnosis.*

Introduction

Ensemble methods are a fairly powerful tool for building a set of machine learning models for the same task to obtain the best indicators of adequacy, accuracy and efficiency of application for data series predictions. In addition, in the absence of «strong learners» to train the model, the use of ensemble methods may be the only correct solution for the task. In this work, studies were carried out on the construction and evaluation of the effectiveness of models by various classification methods for detecting the diagnosis of oncological diseases at early stages [1].

Data collection

Biomarkers can be used to detect disease early, before it becomes clinically evident. Cancer biomarkers are used to detect cancer. Such markers can be obtained on the basis of the so-called tumor microchip – a biochemical analysis based on the principle of forming an array of markers on a microscope slide. Thus, in particular, for the detection of pancreatic cancer, the data of observation of a tumor marker based on the CA-19-9 marker microchip are used [2].

The possibility of presenting and processing the results of serial observations of various markers was studied in the interests of solving the problem of diagnosing oncological diseases. For the experiments, we used a set formed on the basis of data from the early stage of cancer diagnosis, obtained in one of the regional clinics in the Irkutsk region for 10 years.

The studied data set contained the results of observations in 71 cases of oncological diseases of various nature (conditional diagnosis D=1) and for 70 healthy patients from the control group (D=0) (Table 1).

Table 1. *Frequency of occurrence of series of observations of length L*

Таблица 1. *Частота появления серий наблюдений длины L*

D	L										
	1	2	3	4	5	6	7	8	9	10	∑
0	0	2	6	9	4	12	10	9	18	0	70
1	14	7	24	9	10	6	1	0	0	0	71

The data were broken down into series of observations for individual patients. The length L of the series varied from one patient to another. Two types of markers were considered: common and free. There were 683 observations in total, each of which included patient ID, diagnosis, time to final diagnosis, biomarker levels, total and free, and patient age. The classification was based on 4 disease classes (D=1) and a control class (D=0) (Table 2).

Table 2. *Biomarkers data*

Таблица 2. *Данные биомаркеров*

Diagnosis Диагноз	Biomarker A Биомаркер А	Biomarker B Биомаркер В	Biomarker C Биомаркер С
Disease A Болезнь А	5,17	4,17	8,7
Disease B Болезнь В	0,83	0,91	0,88
Disease C Болезнь С	0,75	0,89	1,03
Disease D Болезнь D	0,81	0,86	1,12
D=0	0,89	0,95	0,96

Search for informative signs

To solve the problem of selecting the most informative features, it is advisable to consider the use of various statistical tests and measures of separability of classes. We assumed in our study that the null hypothesis says that the classes presented in Table 2 are inseparable, and the alternative is that the classes are separable.

Statistical tests based on Student's t-Test and Mann–Whitney–Wilcoxon U-Test were aimed at pairwise comparison of two different classes. Therefore, a control class was selected and compared in pairs with disease classes. To evaluate the work of each statistical criterion, p-value is used – the probability of error when the null hypothesis is rejected [3].

For each test, four independent sets of comparisons of each type of diagnosis with the control class were formed.

Several markers with a minimum p-value were selected from each sample. The resulting samples were combined into one set without repeating markers.

Methodology and evaluation of classification efficiency

Data classification was carried out in several stages: applying a sliding control scheme, randomly dividing the initial data sample into training and test data in equal proportions, training the classifier on the training sample and testing it on the test sample, assessing the quality of the classification using one of the described metrics.

A process was created containing classification methods: Logistic Regression, Support Vector Machine, Naive Bayes Classifier, K-Nearest Neighbors, Decision Tree, Random Forest.

In addition, a neural network model was taken separately for the study, trained on the same data and used to compare the experimental results. Model parameter settings (in Python) are shown in Table 3.

Application of the ensemble method

The ensemble learning includes two stages: learning a first-level and a second-level meta-classifier. We use the output of the first-level classifiers as the new features. Next, we use the new features to train the second level meta-classifier (Figure).

Table 3. *Model parameters*

Таблица 3. *Параметры модели*

Model/Модель	Parameter/Параметр
Logistic Regression Логистическая регрессия	max_iter=10, penalty=l2, solver=liblinear,tol=1e-4
Support Vector Machine Метод опорных векторов	decison_function_shape=ovo, C=1,kernel=rbf
Naïve Bayes Наивный байесовский классификатор	alpha=0.01
K-Nearest Neighbors Метод k-ближайших соседей	n_neighbors=10
Decision Tree Дерево решений	max_depth=3, min_samples_leaf=1, criterion=gini
Random Forest Случайный лес	n_estimators=10, max_depth=3, criterion=gini
Ensemble Learning Ансамблевое обучение	layer1=[LR,SVM,NB,KNN,DT,RF], layer2=[LR]

As a result of the analysis, it can be argued that the k-nearest neighbors' method and the random forest method have the least efficiency in diagnosing. At the same time, the naive Bayesian method has the greatest efficiency, although its classification accuracy cannot be considered sufficient. Therefore, it was concluded that the most effective would be to use the ensemble method based on Bayes classifiers [4].

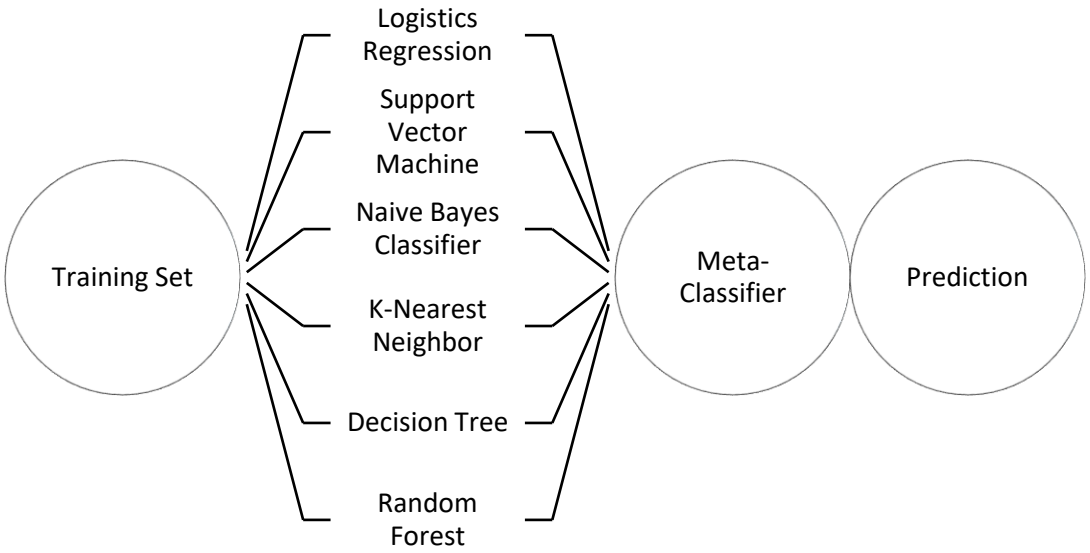

Figure. *Two-level ensemble approach*

Рисунок. *Двухуровневый ансамблевый подход*

Our evaluation method consists of three stages: data processing, model training, and ensemble learning. At the data processing stage, we remove the stop words, punctuation, stemming [5]. At the model training stage, we select logistic regression as the meta-classifier to learn a second-level classifier. Before using ensemble learning, we need to set the hyperparameter of each classifier. At the ensemble learning stage, we use train data and validation data to training each independent classifier, adjust the hyperparameters to achieve the best performance of the independent classifier on the validation set. The experiment results of each model on test data are in Table 4.

Table 4. *Experiment results*

Таблица 4. *Результаты эксперимента*

Model Модель	Accuracy Точность
Logistic Regression/Логистическая регрессия	0,543
Support Vector Machine/Метод опорных векторов	0,432
Naïve Bayes/Наивный байесовский классификатор	0,512
K-Nearest Neighbors/Метод k-ближайших соседей	0,576
Decision Tree/Дерево решений	0,534
Random Forest/Случайный лес	0,489
Ensemble Learning/Ансамблевое обучение	0,681

It can be seen from the table that ensemble learning has achieved the best performance and the performance of the decision tree of a single classifier is the best. Different classifiers can learn different data features, and ensemble learning can integrate the features learned by each classification and the advantages of each classifier. In addition, through the experiments, we found that the performance of the logistic regression and support vector machines is stable, and the classification performance is not obviously different.

Conclusions

The proposed consolidating method for analyzing series of observations based on the fitted model allows one to study any number of markers. The ensemble included not only the listed classifier methods, but also the neural network, which showed comparable accuracy in the previous numerical experiment. The result of the concluded performance vector is very high, i. e. a fairly simple ensemble of methods is able to solve the problem more efficiently than individual methods.

REFERENCES

1. O'Donnell B., Maurer A., Papandreou-Suppappola A., Stafford P. Time-frequency analysis of peptide microarray data: application to brain cancer immunosignatures. *Cancer Inform.*, 2015, vol. 14 (2), pp. 219–233. DOI: 10.4137/CIn.s17285
2. Stafford P., Cichacz Z., Woodbury N.W. Immunosignature system for diagnosis of cancer. *Proc Natl Acad Sci USA*, 2014, vol. 111 (30), pp. E3072–E3080. DOI: 10.1073/pnas.1409432111
3. Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci*, 2008, vol. 3(4), pp. 286–300. DOI: 10.1111/j.1745-6924.2008.00079.x
4. Padhraic S., Wolpert D. Linearly combining density estimators via stacking. *Machine Learning*, 1999, vol. 36, pp. 59–83. Available at: https://doi.org/10.1023/A:1007511322260 (accessed 15 April 2023).
5. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Müller A., Nothman J., Louppe G., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: machine learning in Python. *Journal of machine learning research*, 2011, vol. 12, pp. 2825–2830. Available at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (accessed 21 February 2023).

Received: 19 April 2023.
Reviewed: 10 May 2023.

Information about the authors

Anastasia A. Grigoreva, senior lecturer, Irkutsk National Research Technical University.
Andrei I. Trufanov, Cand. Sc., associate professor, Irkutsk National Research Technical University.
Stanislav V. Grigorev, Cand. Sc., associate professor, Irkutsk National Research Technical University.

УДК 004+616.43

АНСАМБЛЕВЫЕ МЕТОДЫ ДЛЯ РЕШЕНИЯ ЗАДАЧ МЕДИЦИНСКОЙ ДИАГНОСТИКИ

Григорьева Анастасия Александровна[1],
isaevaaa@ex.istu.edu; aag@isti.edu

Труфанов Андрей Иванович[1],
troufan@istu.edu; ait@isti.edu

Григорьев Станислав Валентинович[1],
svg@isti.edu

[1] Иркутский национальный исследовательский технический университет,
Россия, 664074, г. Иркутск, ул. Лермонтова, 83.

Предложен консолидирующий метод анализа рядов наблюдений на основе аппроксимированной модели смеси катализаторов основных компонентов, позволяющий изучать любое количество маркеров. В отличие от лонгитюдного подхода он устраняет необходимость связывать методы регрессионного анализа с их собственными неопределенностями при выборе конкретных моделей. Консолидирующий метод позволяет получить оригинальный результат в предметной области ранней диагностики заболевания: все варианты использования маркеров демонстрируют повышение точности классификации с увеличением продолжительности серии обследований.

Ключевые слова: *ансамблевое обучение, агрегация данных, статистическая классификация, точность модели, медицинский диагноз.*

Информация об авторах

Григорьева А.А., старший преподаватель, Институт информационных технологий и анализа данных, Иркутский национальный исследовательский технический университет.

Труфанов А.И., кандидат физико-математических наук, доцент, Институт информационных технологий и анализа данных, Иркутский национальный исследовательский технический университет.

Григорьев С.В., кандидат химических наук, доцент, Институт информационных технологий и анализа данных, Иркутский национальный исследовательский технический университет.

Поступила 19.04.2023 г.
Принята: 10.05.2023 г.