CIN classification and prediction using machine learning methods

Cite as: AIP Conference Proceedings **1836**, 020010 (2017); https://doi.org/10.1063/1.4981950
Published Online: 05 June 2017

Anastasia Chirkina, Marina Medvedeva, and Evgeny Komotskiy

CIN Classification and Prediction Using Machine Learning Methods

Anastasia Chirkina[1, a)], Marina Medvedeva[1, b)], Evgeny Komotskiy[1, c)]

[1]*Graduate School of Economics and Management, Ural Federal University, Mira, 19, Ekaterinburg, RUSSIA 620002*

[a)]chirkina_nastya@mail.ru
[b)]Corresponding author: marmed55@yandex.ru
[c)]komockye@mail.ru

Abstract. The aim of this paper is a comparison of the existing classification algorithms with different parameters, and selection those ones, which allows solving the problem of primary diagnosis of cervical intraepithelial neoplasia (CIN), as it characterizes the condition of the body in the precancerous stage. The paper describes a feature selection process, as well as selection of the best models for a multiclass classification.

INTRODUCTION

At present, women's health issues are very important not only in medical sphere, but in all socio-economic life of the society. It is especially important nowadays when a trend of significant growth of gynecological diseases in almost all age groups is observed.

Human papilloma virus (HPV) infection is one of the most common infections leading to dysplasia and cervical cancer. In Yekaterinburg, 48.4% of reproductive age women are infected by HPV, 31.8% of which have a cervical disease, and 22.4% have a precancerous state. Cervical intraepithelial neoplasia (CIN) is a severe form of HPV infection, which shows that the present condition of the body is a precancerous one.

The recognition of this disease in the early stages is one of the most important tasks in gynecology.

Early identification of the disease progression helps to prevent the development of cancer and increases the chances of complete recovery and restoration of the female body.

The aim of this work is to develop a model of CIN automatic classification.

The main objectives are:

1. To conduct an analysis of existing data using a variety of decision-making algorithms in the presence of a minimum set of factors;
2. To undertake a reanalysis in case of increasing the number of parameters studied;
3. To analyze and assess the analysis results.

1. METHOD OF DATA COLLECTION

1.1. Data preparation

At the first step, we create a test for our colleagues from the Ural State Medical Academy.

We collect data from 2000+ medical cards of patients and then, using methods of Latent Semantic Analyses, convert anamneses and complaint of patients in groups of questions.

In addition, we use a-priori data and questions from our medical colleagues.

Applied Mathematics and Computer Science
AIP Conf. Proc. 1836, 020010-1–020010-5; doi: 10.1063/1.4981950
Published by AIP Publishing. 978-0-7354-1506-5/$30.00

In the end, we receive a 122-question test that includes the following groups of questions (predictors in the model):

1) Clinical questions (eg. anamnesis, complaint, blood analysis and other);
2) Demographic questions;
3) Socio-psychological questions.

1.2. Data collection

The next step is collection of data to train the model.

With our colleagues from the Ural State Medical Academy we launch open, prospective, cohort controlled study of 508 women, which includes 306 patients with HPV-associated cervical intraepithelial neoplasia I, II, III degree (major groups).

The comparison group (group 4) consistes of 100 patients with HPV-negative cervical intraepithelial neoplasia I-st degree.

The control group (group 5) was represented by 102 women with visually intact cervix.

After all, we receive dataset with the following parameters:

1) Number of items (observations) – 508;
2) Number of input variables – 122;
3) Output variables – 5 classes (CIN I-III, HPV-negative, CIN I-st degree, healthy).

2. MODEL BUILDING

2.1. Feature Selection

It is necessary to carry out the selection of the parameters that have the greatest effect on the classification.

To identify these parameters, the following metrics are used:

1) Pearson correlation;
2) Mutual correlation ;
3) Kendall correlation;
4) Spearman correlation;
5) Chi-square.

After comparison of these methods, the best accuracy was showed when the variables selection using the Chi-square method.

The number of variables in the model, under which the maximum accuracy was achieved, was 10 input variables.

2.2. Model Selection and Evaluation

We split dataset in a ratio of 70% - a training set, and 30% - a test set.

The studies were conducted using Microsoft Azure Machine Learning analytical platform.

When building multiclass models, we used the following algorithm:

1) Multiclass Decision Forest [5];
2) Multiclass Decision Jungles [4];
3) Multiclass Logistic Regression [2];
4) Multiclass Neural Network [1].

There are a lot of methods for classification and data mining (for example [7, 8]), but in this study we used only above-mentioned ones.

The first model was built using the algorithm Multiclass Decision Forest.

The obtained results are shown in Fig. 1. The accuracy of classification - 48.8%.

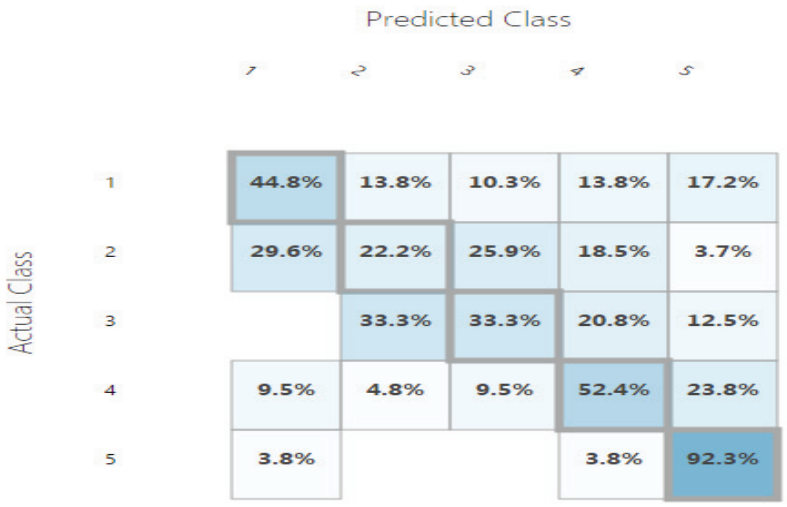

FIGURE 1. Multiclass Decision Forest classification result

Thus, the classification of the fifth group is the best one ("healthy"-class, 92.3%), and the correctness of the fourth group classification - 52.4%. However, 23.8% of fourth group are classified as "healthy", which is unacceptable result.

Multiclass Decision Jungles have the following advantages:

1) Allows combining the branches of a tree;
2) Directed acyclic graph (DAG [3, 6]) takes up less memory and improve performance;
3) The method is stable in the presence of noise characteristics.

The results obtained using Multiclass Decision Jungles are shown in Fig. 2.

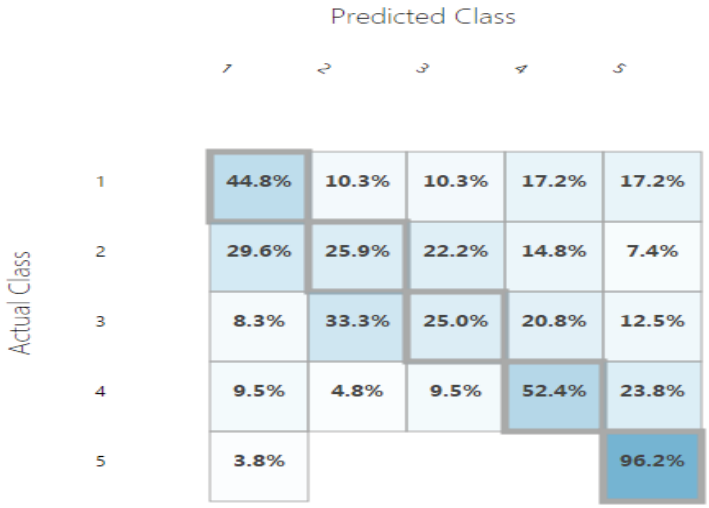

FIGURE 2. Multiclass Decision Jungles classification result

Classification accuracy - 48.8%. The model results are quite similar to those obtained using the Multiclass Decision Forest algorithm. So, one can see that any significant changes of the results did not happen.

The model presented in Fig. 3 was built with the help of Multiclass Logistic Regression.

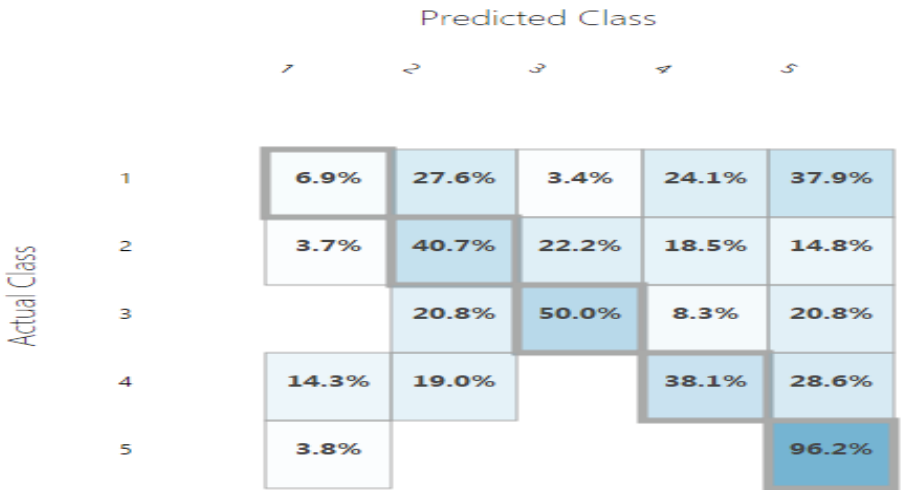

FIGURE 3. Multiclass Logistic Regression classification result

The matrix in Fig. 3 shows that the increased percentage of misclassification is observed for all groups, i.e. the big part of people with the disease is in the "healthy" group, which is unacceptable. The accuracy of this model - 45.6%.

Multiclass Neural Network - is a set of inter-related layers, which converts the input data into the output sequence of weighted edges and nodes. The classification result with Multiclass Neural Network is shown in Fig. 4.

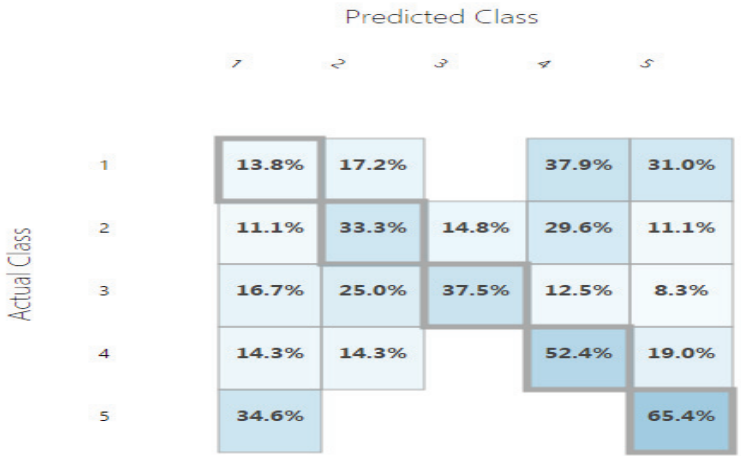

FIGURE 4. Multiclass Neural Network classification result

When using Multiclass Neural Network model accuracy significantly decreased - up to 39.3%. So, here there is also high percentage of incorrect classification of a group of "healthy".

The results of classification accuracy for all used methods are presented in Table 1.

TABLE 1. Algorithms evaluation comparison

Variables number	Accuracy, %			
	Multiclass Decision Forest	Multiclass Decision Jungles	Multiclass Logistic Regression	Multiclass Neural Network
15 variables	48,8	48,8	45,6	39,3
10 variables	54,0	52,0	41,7	43,3

Analyzing the Table 1, one can see an interesting result – the reduction of input variables number improves accuracy of classification for the solving task. Therefore, the best algorithm for CIN detection in our case may be recognized the Multiclass Decision Forest method.

CONCLUSION

The study shows that usage of machine learning techniques allows getting of acceptable model, which is sufficient for separation healthy individuals from patients with CIN. However, at the same time, without using of laboratory methods it is difficult to draw up conclusions about the degree of CIN, so these methods are needed for interpretation of the obtained models.

Nevertheless, the use of machine learning techniques in mobile variant can be useful for a quick pre-self-test, which can determine the risk groups for further more detailed laboratory examination.

ACKNOWLEDGMENTS

Special thanks for our colleague from Ural State Medical Academy and Center protection of motherhood and Infancy Kononova Irina Nikolaevna, without whom this work could not be realized.

REFERENCES

1. P. Kraipeerapun, "Multiclass Classification using Neural Networks and Interval Neutrosophic Sets", Proceedings of the 5th WSEAS Int. Conf. on Computational Intelligence, Man-Machine Systems and Cybernetics (2006).
2. D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression. Second Edition* (Wiley Publishing, Inc., 2000).
3. D. Benbouzid, R. Busa-Fekete, and B. Kegl, "Fast classification using sparse decision DAGs", in *Proceedings of Int. Conf. on Machine Learning (ICML)* (New York, NY, USA, 2012).
4. L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees* (Chapman and Hall/CRC, 1984).
5. L. Breiman, "Random forests", Machine Learning 45(1) (2001).
6. H. Peterson and T. R. Martinez, "Reducing decision trees ensemble size using parallel decision DAGs", Int. Journ. on Artificial Intelligence Tools 18(4) (2009).
7. M.A. Medvedeva and E. I. Komotskiy, "About usage of data mining methods for fraud detection in the sphere of communal services", AIP Conf. Proc. 1738, pp. 110011-1 - 110011-4 (2016); doi:10.1063/1.4951880.
8. Oleg I. Nikonov, Fedor P. Chernavin and Marina A. Medvedeva, "The problems of classification: Method of committees", AIP Conf. Proc., 1738, pp. 110004-1 - 110004-4 (2016). doi:10.1063/1.4951873.