Analysis of variance application in the construction of classifier ensemble based on optimal feature subset for the task of supporting glaucoma diagnosis

Dominika Sułot[1][0000−0002−0246−3223], Paweł Zyblewski[2][0000−0002−4224−6709], and Paweł Ksieniewicz[2][0000−0001−9578−8395]

[1] Department of Biomedical Engineering, Wroclaw University of Science and Technology, Poland
dominika.sulot@pwr.edu.pl
[2] Department of Systems and Computer Networks, Wroclaw University of Science and Technology, Poland
{pawel.zyblewski,pawel.ksieniewicz}@pwr.edu.pl

Abstract. The following work aims to propose a new method of constructing an ensemble of classifiers diversified by the appropriate selection of the problem subspace. The experiments were performed on a numerical dataset in which three groups are present: healthy controls, glaucoma suspects, and glaucoma patients. Overall, it consists of medical records from 211 cases described by 48 features, being the values of biomarkers, collected at the time of glaucoma diagnosis. To avoid the risk of losing information hidden in the features, the proposed method – for each base classifier – draws a separate subset of the features from the available pool, according to the probability determined by the ANOVA test. The method was validated with four base classifiers and various subspace sizes, and compared with existing feature selection methods. For all of the presented base classifiers, the method achieved superior results in comparison with the others. A high generalization power is maintained for different subspace sizes which also reduces the need to optimize method hyperparameters. Experiments confirmed the effectiveness of the proposed method to create an ensemble of classifiers for small, high-dimensional datasets.

Keywords: Analysis of variance · Glaucoma classification · Subspace selection · Non-uniform random subspace · Classifier ensembles

1 Introduction

Sight is the main and most important human sense. Using its capabilities, we are receiving most of the information coming to us from the surrounding world. Therefore, the narrowing of the field of view – often irrevocably leading to total blindness – make it much more difficult to perform everyday activities. One of the main factors leading to such disorders is glaucoma. It is estimated that over

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_10

70 million people are suffering from this disease, but due to its asymptomatic course, even over 50% of those affected may not be aware of it [15]. The most common type of glaucoma is *primary open-angle glaucoma*, which does not give any symptoms until the later stages, and its effects are irreversible. This form of the disease will be taken into consideration in this article and will be later abbreviated as *glaucoma*.

The current research is focusing on the early detection of glaucoma, at the stage preceding irreversible changes – especially – visual field narrowing [4, 10, 9]. Biomarkers are the most commonly used for this purpose. They include, among others, *intraocular pressure* (IOP), *retinal nerve fiber layer* (RNFL) *thickness*, parameters concerning the position and shape of the *lamina cribrosa*, size and shape of the optic nerve disc and many others. Their observation and analysis can have an impact on the early detection of developing glaucoma and can help in trying to control the disease [2].

The dynamic development of machine learning methods allows to support medical diagnostics [3, 7]. However, due to a large number of available biomarkers in the case of glaucoma, automatic classification of medical cases into disease groups, using pattern recognition methods, states a non-trivial problem. It results from limited datasets and a large number of features describing each analyzed sample. This problem is widely known as a *curse of dimensionality* [1]. There are different methods to cope with this problem [11], mainly based on the feature selection, i.e. selecting the best subset of the available feature space, which will be used for further analysis. However, this solution is not always effective, and by rejecting a certain number of features, the information contained in them is lost. Additionally, sometimes long-term optimization of parameters is needed to obtain a satisfactory result. An example of such a technique may be an application of statistical methods to rank the features [5] or to use *Principal Component Analysis* to reduce the size of the dataset [16].

The development of these methods and, at the same time, the solution to the problem of rejecting input features is the use of *ensemble learning* to train classifiers based on different subspaces. The most common type of such processing is using the *Random Subspace* method, also known as feature bagging [6]. It consists in drawing and returning features for a separate subspace for a single classifier.

The following paper proposes a novel method that includes the basics of the two above-mentioned methods. It allows to automatically build an ensemble of classifiers on a non-randomly drawn subspace of features, where the probability of drawing depends on the ranking obtained with the use of *Analysis of Variance* (ANOVA) [14]. Thus, each classifier is learned from a smaller amount of data, avoiding *the curse of dimensionality*, and at the same time, no features are rejected, minimizing the risk of rejecting valuable information contained in them. Such an approach leads to an increase in the overall diversity of the trained pool and allows to achieve high classification quality. The proposed method was named ANOVA *Subspace Ensemble* (ANOVA SSE). Finally, the proposed method was tested on a numerical medical dataset, and a built ensemble was able to classify glaucoma progression groups.

The main contributions of this study are:

- a novel solution to diversify a homogeneous pool of classifiers based on the analysis of variance,
- experimental evaluation of the method in the context of standard solutions to the problem,
- statistical analysis of the obtained results.

2 Methods

The paper focuses on a method, that aims to extract a set of feature subsets from a dataset and at the same time, creating a classifier ensemble, in which each classifier will be trained on a different subset. The method is based on the ANOVA test, used to calculate the probability with which a given feature will be drawn from the entire set of features. From the results of ANOVA, F-value for each feature is taken, and transformed in the way, that the sum of an array created from this set of F-values will be equal to one. This operation is performed so that the F-value vector may be interpreted as a discrete probability distribution.

The created array is passed to a function that generates the random sample from a features vector, as a probability for each entry. Finally, the function, according to the given probability, will draw from the set of features, increasing the probability of drawing the features that obtained the greater F-value. Then, as many subsets as there are classifiers in the ensemble are drawn, and each classifier is trained on a separate subset.

Further, in the training process, weight is calculated for each classifier in the ensemble based on its *balanced accuracy* on the training set. The aforementioned procedure, which is the classifier fit function, is described in the form of pseudocode in Algorithm 1. At the final prediction, the supports of each are multiplied by the weights and then the standard accumulation of the supports is done. The method uses the information contained in all the features, not rejecting any of them, while favoring features that have higher F-value, thus maximizing the quality of the created ensemble.

3 Dataset

The dataset used in this study is a retrospective data collection, described in more detail in [8]. It consists of a set of biomarker values for each of the patients. These values are typically acquired during glaucoma diagnosis and are commonly used. The set includes intraocular pressure (IOP), retinal nerve fiber layer thickness (RNFL), optic disk morphology parameters, and many others. An experienced ophthalmologist assigned each patient individually to one of three groups: healthy controls, glaucoma suspects, and glaucoma patients, based on the collected data and images acquired with optical coherence tomography. The entire collection contains data from 211 patients (69 controls, 72 glaucoma suspects,

Algorithm 1: The fit function for the ANOVA SSE method

```
Input: $X$ as an array of training examples with $y$ containing corresponding
       labels and $n$ as an subspace size and $k$ as a size of ensemble
Output: A list of trained base classifier (ensemble) and corresponding weights
n_features := number of features avaiable in $X$;
$p$ := the list of F-value for each feature in $X$ obtained from ANOVA;
$p$ := p/sum($p$);
$f$ := [0, 1, ..., n_features];
classfiers := a list of $k$ base classifiers;
for classifier in classifiers do
    ss_indexes := a list of $n$ drawn numbers from $f$ with the probability of $p$;
    train classifier on all samples from $X$ but only on features with ss_indexes;
    append trained classifier to a list of ensemble;
    append weight, calculated as a balance accuracy score for classifier
      determined on those training samples, to a weights list ;
end
```

70 glaucoma patients), and for each of them, there are 48 features available. Each patient from whom the data was derived gave their written consent and the studies were approved by the *Bioethical Committee of the Wroclaw Medical University (KB–*332/2015*)*.

4 Experiment design

The whole experiment was conducted using *Python* language and *scikit-learn 0.23.2* [12] package. The implementation of both the proposed method and experimental code, to preserve the possibility of replication of performed experiments, is publicly available in Github repository[1].

The performance evaluation, as well as the comparison of the various methods, was based on the *balanced accuracy* score, which is calculated as the arithmetic mean of specificity and sensitivity. The *t-test with non-parametric correction* was used to check whether the results obtained with the different methods are statistically dependent [13]. The 5x5 repeated cross-validation protocol was used to obtain reliable results, both for the proposed and reference methods.

Four base classifiers were used and validated in the experiments: *Multilayer Perceptron* (MLP), *k-Nearest Neighbors* (kNN), *Classification and Regression Trees* CART and *Support Vector Machines* classifier SVC. The parameters of the individual classifiers used for the experiments are presented in Table 1.

For comparison, models based on the two most common feature selection methods were calculated, i.e. *Random Subspace* and the method of selecting only the k most differentiable features based on the ANOVA test (*k-best*). In addition, the results for simple, single models, that were built on a full available feature space, are also presented.

[1] https://github.com/w4k2/anova_sse

Table 1. The parameters of the classifiers that were set during the computational experiments.

Classifier	Parameters
MLP	hidden layers = 100; activation = the rectified linear activation function; solver = Adam; alpha = 0.0001; constant learning rate = 0.001; maximum number of iterations = 20; beta 1 = 0.9; beta 2 = 0.999; epsilon = 1e-8
kNN	number of neighbours = 5; uniform weights; leaf size = 30; Euclidean metric
DTC	Gini impurity criterion; without maximum depth; the minimum number of samples required to split an internal node = 2; the minimum number of samples required to be at a leaf node = 1
SVC	linear kernel; with the enabled probability estimates; number of iteration = 1; one vs rest decision function shape

The aim of the experiments was to verify the effectiveness of the proposed method both as a method for selecting a subspace on which classifier ensemble was to be trained, as well as a method operating on a small part of the feature subspace from the initial data set (e.g. due to the acceleration of learning). Therefore, for solutions based on feature selection, experiments were performed for several sizes of subspace ranging from 1 to 48.

5 Experimental evaluation

The obtained results are shown in Table 2 and Figure 1. They are the mean values and standard deviations calculated across folds. Additionally, the approaches for which the given method is statistically better are marked under the results in the table. What may be observed, with a size of feature subspace greater than 18, none of the considered methods is statistically better than the proposed one. Comparing them using the same base classifier leads to the observation that the proposed method receives statistically better or no worse results than the other two feature-selection methods.

Furthermore, as can be seen in Figure 1, it always get better results than the base solution trained on the all features, regardless of the number of attributes used. By analyzing plots, it can also be concluded that the obtained balanced accuracy maintains its stability even for a large number of features. Which means that a long-term optimization of the subspace size is not needed to get a good result, unlike the method based on the selection of k-best features.

Additionally, in cases where random subspace achieves high balanced accuracy, the proposed method is also able to achieve a similar level of accuracy, but for a much smaller number of features. The graphs also show a black, dashed vertical line, which is the place where the number of features is equal to the root of all available features. Here our method in each case achieves better results than random subspace.

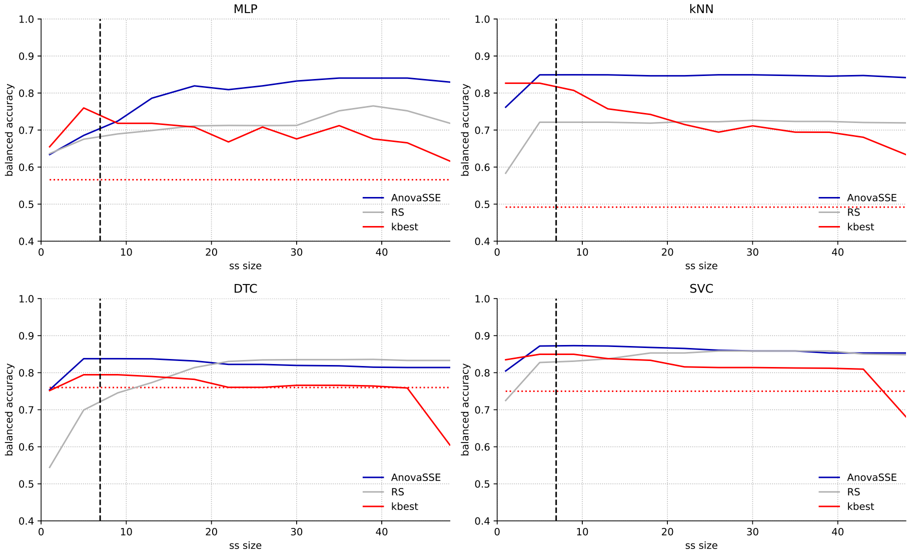

Fig. 1. The dependence of the balanced accuracy on the size of the feature subspace for various methods based on the selection of features and for the classifiers learned on the whole set of features (dotted red). Additionally, a black vertical line marks the number of features corresponding to the square root of all available in the set.

Summarizing the results presented in Table 2, the proposed method shows the potential for being used to create an ensemble of classifiers based on different subspaces of features, while the size of these subspaces is not critical and does not significantly affect the results. Ultimately, the maximum score obtained with this method is .872 using SVC as a base classifier and feature subspace size of 9. This model is statistically superior to almost any model based on the same subspace size.

The results averaged over all four base classifiers are presented on the left side of Table 3. They show that the proposed method always obtained statistically better results than the other two, with the size of the feature subspace greater than 9. In the case of size 9, there is no statistical difference between the proposed one and the method consisting in using 9 best features based on the ANOVA test. From the obtained results it can be concluded that for the considered problem of glaucoma progression group classification the best choice of a base classifier is SVC, which is always statistically better or not worse than the other achieving the highest result of averaged balanced accuracy up to .843.

6 Conclusions

This paper takes up the topic of generating an ensemble of classifiers on the basis of a high-dimensional dataset. The proposed method tries to solve this problem

Table 2. The mean value and the standard deviation for all of the considered methods based of feature selection for four different base classifiers. The results are presented for different sizes of the used feature subspace. The numbers of the methods for which the method obtained statistically better results are also shown under the results.
ss size is an abbreviation for the utilized size of subspace.

ss	K-BEST				RANDOM SUBSPACE				ANOVA SSE			
size	MLP (1)	kNN (2)	DTC (3)	SVC (4)	MLP (5)	kNN (6)	DTC (7)	SVC (8)	MLP (9)	kNN (10)	DTC (11)	SVC (12)
9	.781 ±.062 5,9	.807 ±.056 5, 6, 9	.794 ±.052 5, 9	.850 ±.057 3, 5–7, 9	.675 ±.048 —	.721 ±.063 —	.746 ±.065 —	.831 ±.059 5–7, 9	.686 ±.029 —	.849 ±.064 5–7, 9	.836 ±.054 5–7, 9	.872 ±.055 1–3, 5–9
18	.668 ±.061 —	.742 ±.065 —	.782 ±.059 1	.833 ±.055 1, 2, 5, 6	.736 ±.048 1	.723 ±.063 —	.814 ±.062 1, 5, 6	.853 ±.060 1–3, 5, 6, 9	.786 ±.046 1, 2	.847 ±.055 1, 2, 5, 6, 9	.832 .058 1–3, 5, 6	.868 ±.055 1–3, 5, 6, 9
30	.717 ±.068 —	.694 ±.066 —	.766 ±.058 —	.813 ±.049 1, 2, 5, 6	.712 ±.070 —	.723 ±.073 —	.835 ±.055 1, 2, 5, 6	.859 ±.056 1–3, 5, 6	.832 ±.052 1, 2, 5, 6	.849 ±.050 1–3, 5, 6	.819 ±.062 2, 5, 6	.861 ±.060 1–3, 5, 6
39	.712 ±.063 —	.694 ±.056 —	.764 ±.061 2	.810 ±.059 1, 2, 6	.752 ±.059 —	.723 ±.071 —	.830 ±.060 1, 2, 5, 6	.862 ±.058 1–3, 5, 6	.840 ±.061 1, 2, 5, 6	.847 ±.056 1, 2, 5, 6	.814 ±.060 1, 2	.853 ±.068 1–3, 5, 6
48	.616 ±.043 —	.634 ±.061 —	.605 ±.055 —	.681 ±.066 1, 3	.719 ±.042 1–3	.721 ±.066 1, 3	.833 ±.054 1–6	.849 ±.054 1–6	.846 ±.052 1–6	.852 ±.052 1–6	.814 ±.059 1–6	.861 ±.069 1–6

with the use of all available features, so as not to reject important information that is hidden in these features that less differentiating classes. The method is presented on a demanding dataset, which is above all small but also contains many features that define each object. This set includes three classes: healthy controls, glaucoma suspects, and glaucoma patients. The proposed method shows that, in comparison with other methods based also on feature selection, it can achieve the highest results, which was confirmed by statistical tests that further support the benefits of using non-uniform feature selection. An additional advantage is that the method is characterized by only small fluctuations in balanced accuracy when changing the size of the feature subspace. This reduces the need for a time-consuming process to search parameters to find the optimal size. Additionally, in solutions where the random subspace method turns out to be effective, the proposed method allows achieving similar results of accuracy with a much smaller size of the feature subspace, speeding up the learning process.

Acknowledgments

The work of D. Sułot was supported by *InterDok – Interdisciplinary Doctoral Studies Projects* at *Wroclaw University of Science and Technology*, a project co-financed by the European Union under the European Social Fund, while P. Ksie-

Table 3. The mean value and the standard deviation for all of the considered methods averaged over base classifiers (left side) and of the considered base classifiers averaged over methods (right side). The results are presented for different sizes of the used feature subspace. The numbers/letters of the methods for which the method obtained statistically better results are also shown under the results.
ss size is an abbreviation for the utilized size of subspace.

ss size	K-BEST	RANDOM SUBSPACE	ANOVA SSE	MLP	kNN	DTC	SVC
	(1)	(2)	(3)	(a)	(b)	(c)	(d)
9	.808 ±.044 2	.743 ±.048 —	.811 ±.042 2	.714 ±.033 —	.793 ±.051 a	.793 ±.044 a	.851 ±.051 a, b, c
18	.756 ±.044 —	.781 ±.043 —	.833 ±.043 1, 2	.730 ±.035 —	.771 ±.048 —	.809 ±.050 a, b	.852 ±.050 a, b
30	.748 ±.034 —	.782 ±.046 —	.840 ±.046 1, 2	.754 .044 —	.755 ±.050 —	.803 ±.049 —	.841 ±.053 a, b
39	.745 ±.038 —	.792 ±.050 1	.839 ±.049 1, 2	.768 ±.051 —	.755 ±.050 —	.803 ±.049 —	.841 ±.053 a, b
48	.634 ±.048 —	.780 ±.041 1	.843 ±.047 1, 2	.727 ±.034 —	.735 ±.044 —	.751 ±.042 —	.797 ±.048 a, b, c

niewicz and P. Zyblewski were supported by the *Polish National Science Centre* under the grant No. 2017/27/B/ST6/01325.

References

1. Bellman, R.: Curse of dimensionality. Adaptive control processes: a guided tour. Princeton, NJ **3**, 2 (1961)
2. Beykin, G., Norcia, A.M., Srinivasan, V.J., Dubra, A., Goldberg, J.L.: Discovery and clinical translation of novel glaucoma biomarkers. Progress in Retinal and Eye Research p. 100875 (2020)
3. Goecks, J., Jalili, V., Heiser, L.M., Gray, J.W.: How machine learning will transform biomedicine. Cell **181**(1), 92–101 (2020)
4. Gupta, K., Thakur, A., Goldbaum, M., Yousefi, S.: Glaucoma precognition: Recognizing preclinical visual functional signs of glaucoma. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020)
5. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research **3**(Mar), 1157–1182 (2003)
6. Ho, T.K.: The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence **20**(8), 832–844 (1998)
7. Jackowski, K., Jankowski, D., Ksieniewicz, P., Simić, D., Simić, S., Woźniak, M.: Ensemble classifier systems for headache diagnosis. In: Information Technologies in Biomedicine, Volume 4, pp. 273–284. Springer (2014)

8. Krzyżanowska-Berkowska, P., Czajor, K., Robert, I.D.: Associating the biomarkers of ocular blood flow with lamina cribrosa parameters in normotensive glaucoma suspects. comparison to glaucoma patients and healthy controls. PLoS One (2021 (in review))
9. Krzyżanowska-Berkowska, P., Czajor, K., Syga, P., Iskander, D.R.: Lamina cribrosa depth and shape in glaucoma suspects. comparison to glaucoma patients and healthy controls. Current eye research **44**(9), 1026–1033 (2019)
10. Kurysheva, N.I., Parshunina, O.A., Shatalova, E.O., Kiseleva, T.N., Lagutin, M.B., Fomin, A.V.: Value of structural and hemodynamic parameters for the early detection of primary open-angle glaucoma. Current Eye Research **42**(3), 411–417 (2017)
11. Mwangi, B., Tian, T.S., Soares, J.C.: A review of feature reduction techniques in neuroimaging. Neuroinformatics **12**(2), 229–244 (2014)
12. Pedregosa, F., et al: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
13. Santafe, G., Inza, I., Lozano, J.A.: Dealing with the evaluation of supervised classification algorithms. Artificial Intelligence Review **44**(4), 467–508 (2015)
14. Tabachnick, B.G., Fidell, L.S.: Experimental designs using ANOVA. Thomson/Brooks/Cole Belmont, CA (2007)
15. Weinreb, R.N., Aung, T., Medeiros, F.A.: The pathophysiology and treatment of glaucoma: a review. Jama **311**(18), 1901–1911 (2014)
16. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)