Feature Selection Based on Mrmr and Genetic Algorithm for Data Classification Problems

Maksim Tislenko
Samara National Research University
Samara, Russia
e-mail: makstislenko@gmail.com

Rustam Paringer
Samara National Research University
Samara, Russia
e-mail: rusparinger@ssau.ru

Alexander Kupriyanov
Samara National Research University
Samara, Russia
e-mail: akupr@ssau.ru

Dmitriy Kirsh
Samara National Research University
Samara, Russia
e-mail: limitk@mail.ru

Artem Mukhin
Samara National Research University
Samara, Russia
e-mail: mukhin.av@ssau.ru

David Asatryan
Institute for Informatics and
Automation Problems of NAS RA
Yerevan, Armenia
e-mail: dasat@iiap.sci.am

Mariam Haroutunian
Institute for Informatics and
Automation Problems of NAS RA
Yerevan, Armenia
e-mail: armar@sci.am

***Abstract*—This paper proposes a new feature selection algorithm based on genetic algorithms and the MRMR algorithm. The algorithm aims to select a subset of features that maximizes classification accuracy. The algorithm works by creating a population of subsets of features, evaluating their weighted F1-score, and then hybridizing the best individuals to create new subsets. The probability of adding a feature during hybridization is estimated using the MRMR algorithm. Finally, the best subset of features is selected based on the F1-score. The proposed algorithm is compared to existing random forest and univariate feature selection algorithms. Three datasets were used to compare them. The proposed algorithm showed better accuracy on average by 0.015 on these three datasets. The proposed algorithm has the potential to improve classification accuracy in datasets where existing feature selection algorithms are insufficient.**

***Keywords*—Feature selection, genetic algorithm, MRMR algorithm, univariate feature selection, random forest algorithm, weighted F1-score.**

I. INTRODUCTION

Feature selection showed good results in pre-processing data for building a model in machine learning. It allows us to speed up data classification, select a subset of features in which the relationship between properties and the target value is easier to trace, and remove unnecessary features. Choosing such a subset of features that the classification accuracy on it will be maximum among all subsets of the same cardinality is a main goal of feature selection during data preprocessing.

The purpose of this work is to create our own feature selection algorithm for data classification problems, which may allow us to classify objects with greater accuracy compared to the existing random forest and univariate feature selection algorithms. It may be necessary for some of the datasets where the existing feature selection algorithms used in preprocessing can’t allow getting enough classification accuracy.

II. ALGORITHM DESCRIPTION AND COMPARISON

As a solution, it is proposed to use a feature selection algorithm based on a genetic algorithm, in which the probability of adding a feature during hybridization is estimated using the MRMR (maximum relevance minimum redundancy) algorithm. The essence of MRMR algorithm is that we strive to choose a subset of features that has the maximum correlation (relevance) with the target value, and the features inside this subset have minimal correlation (redundancy) with each other. These two algorithms were chosen because of their good performance on feature selection. It is shown in [1] that MRMR algorithm shows one of the best results among algorithms that have been compared and in [2] demonstrated that the genetic algorithm shows results with the same quality of feature selection.

The main idea of a genetic algorithm in feature selection is to create a population, feature subsets, select parents, pairs of subsets, crossover, create a child subset consisting of features contained in the parents, and mutation, modification of the child subset by adding features that are not contained in parents. All of that resemble the biological process of chromosome formation.

The following is a step-by-step selection algorithm.

1. In the first step, a population is created from various sets of features. The population size is equal to the number of features in them. In [3], it is shown that this number is optimal. The number of selected features in each subset is an order of magnitude less than in the original dataset on which the classification is made. The features for the individuals in the initial population are chosen at random.

https://doi.org/10.51408/csit2023_10

2. In addition, the feature quality is precomputed and is based on univariate feature selection algorithm, and the metric for this algorithm is selected as a hyperparameter. The feature quality score is stored as a sorted array.
3. Further in the cycle, until the quality of the best individual from the population begins to stagnate, the individuals are hybridized. Stagnation is understood as the lack of an increase of the maximal weighted F1-score that is evaluated over all individuals over three iterations.
3.1. To do this, the quality of each individual in the population is evaluated using a weighted F1 score.
3.2. After this, pairs of individuals which will be hybridized are selected. The number of children is equal to the number of parents in the population. Parents with a better F1-weighted score are more likely to be used for hybridization.
3.3. Before hybridization, a mutation occurs: from the set of those features that are not in the parents, several features are selected to be added to the child. The probability of adding a feature depends on the quality in the set of features calculated at the beginning. The number of features added depends on the number of identical features in each of the two crossed parents, as well as on the value of the weighted F1-score for each of the parents. The more such features and the smaller value of the weighted F1-score means that more features that are not contained in two parents will be added to the child.
3.4. Next comes the addition of features from the parents, the probability of adding is calculated based on the MRMR algorithm on the child.
4. After leaving the cycle, when stagnation occurs, an individual from the population is taken, which shows the best result in classification, this subset of features will be returned as a result.
In Fig. 1 the scheme of the proposed feature selection algorithm is shown.

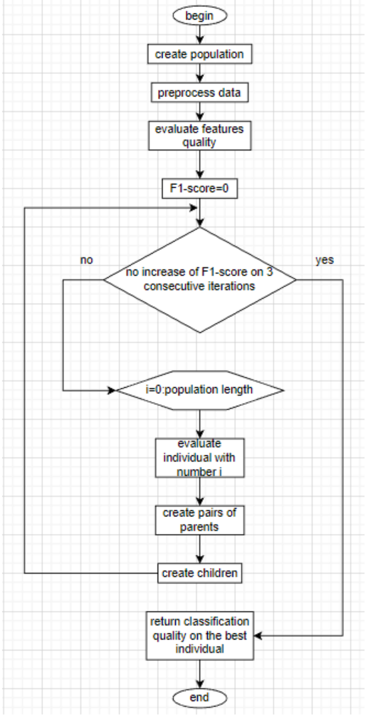
begin
create population
preprocess data
evaluate features quality
F1-score=0
no
yes
no increase of F1-score on 3 consecutive iterations
i=0:population length
evaluate individual with number i
create pairs of parents
create children
return classification quality on the best individual
end

Fig. 1. Scheme of algorithm

The results of the algorithm were compared with the results of the univariate feature selection and random forest algorithms from the Scikit-learn library on The broken machine [4] dataset, Breast cancer dataset [5] and Wine dataset [6]. The tables below show the results of each feature selection algorithm on datasets and corresponding classifiers. In the diagrams, the ordinate shows the values of the weighted average of the F1-score of the classification results for each of the data sets using different selection methods and without them, the abscissa shows the compared algorithms.
The broken machine dataset deals with manufacturing equipment data. There are 58 different unnamed features and 900000 observations. This dataset can be used to create a classification model for predicting breakdowns. Thus, there are 2 classes – the machine is broken or not. Missing values in columns were replaced by column averages.

Table 1. Weighted F1-scores for different algorithms on The broken machine dataset

	Logistic regression	Random Forest
Chi-squared	0,508	0,575
Mutual information	**0,526**	0,570
F–statistics in ANOVA	0,520	0,616
Random Forest	0,520	0,586
Proposed algorithm	0,516	**0,624**

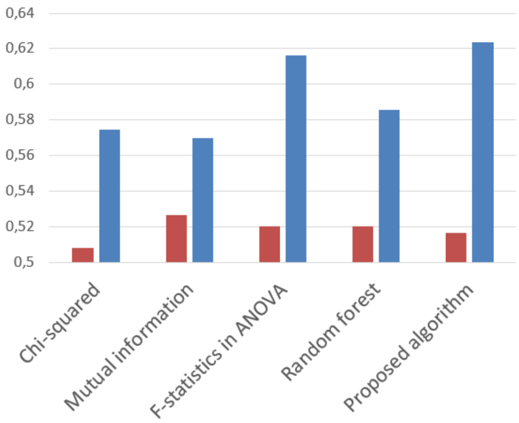

Fig. 2. Weighted F1-scores for different algorithms on the broken machine dataset

The diagram shows that the best classification using logistic regression occurs when feature selection is performed using mutual information. It is also worth noting that the proposed algorithm shows a result worse than the univariate feature selection algorithm by 0.0101. The result is better than we use the Chi-square criterion in the univariate feature selection algorithm by 0.0083. The diagram also shows that the proposed algorithm, when we use a classifier based on a random forest, has better results than any of the existing algorithms. The value of the weighted F1-score when we use the proposed algorithm on 0.0078 more than when using the best of the compared existing algorithms, algorithm univariate feature selection with the metric F-statistic in ANOVA. In general, it is worth noting that for different classifiers, different feature selection algorithms may be optimal. However, on average for two classifiers, the value of the weighted F1-score when we use the proposed algorithm was 0.0017 more than on the best existing algorithm in this indicator, the algorithm univariate feature selection with the F-statistics in ANOVA, using which the average score is 0.5682.

Table 2. Weighted F1-scores for different algorithms on the Breast cancer dataset

	Logistic regression	Random Forest
Chi-squared	0,926	0,924
Mutual information	0,930	0,913
F–statistics in ANOVA	0,916	0,930
Random Forest	0,934	0,933
Proposed algorithm	**0,935**	**0,935**

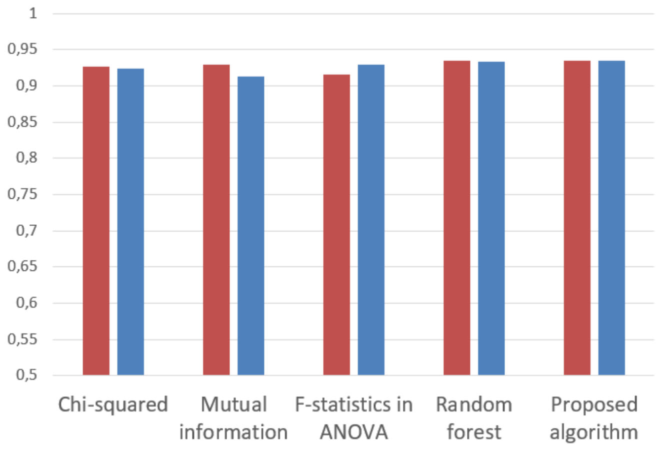

Fig. 3. Weighted F1-scores for different algorithms on the Breast cancer dataset

The diagram shows that the highest value of the weighted F1-score, both using logistic regression and using a random forest as a classifier, is achieved when we use the proposed algorithm. Among the existing feature selection algorithms, the random forest algorithm showed the best results. It should be noted that when we use logistic regression, the values of the weighted F1-score in the selection of features by the proposed algorithm and the random forest algorithm differ only by 0.0003, while when using the random forest algorithm as a classifier, the difference increases to 0.0017.

Among the algorithms, univariate feature selection for a classifier based on a random forest, the largest value of the weighted F1-score is achieved when we use F-statistics in ANOVA, for logistic regression - when we use mutual information metric. However, the value of the weighted F1-score when we use the univariate feature selection algorithm with any metric is worse by more than 0.005 compared to the proposed algorithm.

Table 3. Weighted F1-scores for different algorithms on Wine dataset

	Logistic regression	Random Forest
Chi-squared	0,911	0,955
Mutual information	0,934	**0,961**
F–statistics in ANOVA	0,933	0,950
Random Forest	**0,945**	0,955
Proposed algorithm	0,934	0,960

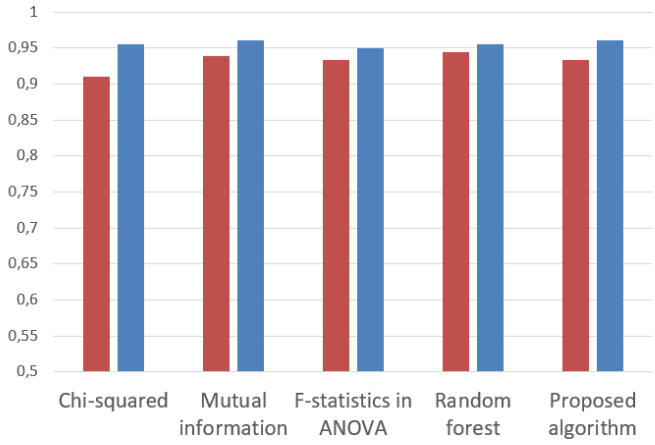

Fig. 4. Weighted F1-scores for different algorithms on the Wine dataset

The diagram shows that the highest value of the weighted F1-score for the classifier based on logistic regression is achieved when we use the feature selection algorithm based on random forest. For the proposed algorithm, the result is less by 0.0106. It is also worth noting that in addition to the random forest algorithm, the value of the weighted F1–score was greater than on the proposed algorithm when we use the univariate feature selection algorithm with the mutual information metric.

When we use the proposed algorithm together with a Random forest classifier, the results are almost the same as when we use a univariate feature selection algorithm with mutual information metric, the difference is 0.0003. The third result was shown by Random forest algorithm, however, its value of the weighted F1–score is less than that of the proposed algorithm by 0.0053.

On the Wine dataset, the indicator of the weighted F1-score of the proposed algorithm is less than the univariate feature selection algorithm with the mutual information metric. This may be due to the fact that there are a small number of observations in the Wine dataset compared to The Broken Machine and Breast cancer dataset, and that's why, the genetic algorithm can quickly converge to a local optimum.

III. Conclusion

Through computational experiments on three datasets, it was discovered that the proposed algorithm, employing a Random Forest classifier, achieves an average increase of 0.015 in the weighted F1-score compared to the feature selection algorithm based on Random Forest. Notably, the proposed feature selection algorithm was identified as the best-performing algorithm among the options considered for these specific datasets. By employing the described algorithm, along with appropriate classifiers, improved data classification accuracy can be increased compared to using univariate feature selection and random forest algorithms available in the Scikit-learn library. This approach can be applied to preprocess datasets from diverse industries. The described algorithm was implemented as a software package and it uses a programming interface corresponding to the Scikit-learn library, that’s why it can be easily used for data preprocessing [7].

ACKNOWLEDGMENT

This work was supported by the Russian Foundation for Basic Research and RA Science Committee within the framework of the joint research project RFBR 20-51-05008 Arm_a and SCS 20RF-144, accordingly.

REFERENCES

[1] V. Vora and H. Yang, “Comprehensive Study of Eleven Feature Selection Algorithms and their Impact on Text Classification”, *2017 Computing Conference*, London, England, pp. 440–449, 2017.

[2] A. El Akadi, A. Amine, A. El Ouardighi and D. Aboutajdine, "A New gene selection approach based on Minimum Redundancy-Maximum Relevance (MRMR) and Genetic Algorithm (GA)," *2009 IEEE/ACS International Conference on Computer Systems and Applications*, Rabat, Morocco, 2009, pp. 69-75, doi: 10.1109/AICCSA.2009.5069306.

[3] J. Alander,“On optimal population size of genetic algorithm”, *CompEuro 1992 Proceedings Computer Systems and Software Engineering*, The Hague, pp. 65-70, 1992.

[4] The broken machine [Electronic resource]. — Access mode: https://www.kaggle.com/ivanloginov/the-broken-machine (05.02.2022)

[5] Sklearn.datasets.load_breast_cancer [Electronic resource]. — Access mode:https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html (15.05.2023).

[6] Sklearn.datasets.load_wine [Electronic resource]. — Access mode: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html (15.05.2023).

[7] genmrmr 0.1.0 [Electronic resource]. — Access mode: https://pypi.org/project/genmrmr/ (15.05.2023).