Comparison of feature selection algorithms for data classification problems

M. D. Tislenko
Samara National Research University
Samara, Russia
makstislenko@gmail.com

A. V. Gaidel
Samara National Research University
Samara, Russia
andrey.gaidel@gmail.com

Abstract

This article discusses various feature selection algorithms, compares the classification accuracy with feature selection algorithms and without them on different datasets.

Introduction

Feature selection, as a data preprocessing strategy, has been proven to be effective and efficient in preparing data (especially high-dimensional data) for various data-mining and machine-learning problems. The objectives of feature selection include building simpler and more comprehensible models, improving data-mining performance, and preparing clean, understandable data[1]. There are a huge number of different methods to determine the best subset of features. This problem is NP-hard, the guaranteed optimal solution can only be found by exhaustive enumeration, which can take a long time for a large number of attributes[2].

Feature selection should be distinguished from feature extraction. Although, both techniques are used to reduce the number of features in a dataset, feature extraction is reduction technique in dimensionality that creates new combinations of attributes, whereas feature selection includes and excludes the attributes that are present in the data without changing them[3].

The objective of this research is to compare different methods of feature selection to improve the choice of algorithm for solving data classification problems.

Materials and methods

To compare different methods of feature selection, 4 datasets were taken.

In [6], the CSV file contains 5172 rows, each row for each email. There are 3002 columns in total. The last column has predictive labels: 1 for spam, 0 for non-spam. The remaining 3000 columns are the 3000 most common words in all emails after excluding non-letter characters/words. For each row, the count of each word (column) in that email (row) is stored in the corresponding cells.

In [7], the data used in this set were collected from 188 Parkinson's disease patients (107 males and 81 females) aged 33 to 87 years. After examination by a physician, a sustained sounding of the vowel a with three repetitions was obtained from each subject. As attributes, the results of various sound processing algorithms were taken, a total of 755 attributes.

In [8], the data set concerns the accumulated data on production equipment. There are 58 different indicators, and they are all unnamed. So, the purpose of this dataset is to create a classification model for predicting breakdowns.

In [9], the dataset contains 225 financial figures that are typically found in the tens of thousands of filings published each year by every public trading company in the US to sell shares.

All datasets had missing values. These missing values have been replaced with the mean values of this feature.

In this research Logistic Regression and Random Forest for classification with SelectKBest and RandomForest algorithms for feature selection were used as mentioned above. Chi-squared, information gain and F-score of ANOVA were used as statistical criterions for Select K Best algorithms.The StandardScaler from the scikit-learn library was used for data preprocessing.

Experimental results and discussion

The tables below show the results of each feature selection algorithm on datasets [6]-[9] and corresponding classifiers. In the diagrams, the ordinate shows the values of the weighted average of the F-measure of the classification results for each of the data sets using different selection methods and without them, the abscissa shows the data sets on which the classification is carried out.

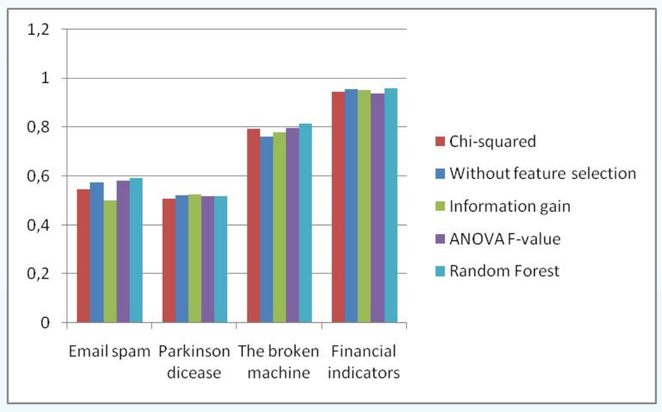

Fig. 1. Diagram of classification using logistic regression

TABLE I. Results of classification using logistic regression

	Email spam	Parkinson disease	The broken machine	Financial indicators
Without feature selection	0.5751	0.5206	0.7617	0.9562
Chi-squared	0.5469	0.5078	0.7931	0.9474
F-statistics ANOVA	0.5837	0.5202	0.7961	0.9403
RandomForest	0.5934	0.5199	0.8146	0.9595
Information gain	0.5	0.5262	0.7816	0.952

The diagram shows that the best classification using logistic regression occurs when feature selection carrying out using a random forest, the proportion of correctly classified objects is approximately 2% higher than using classifying without feature selection, and on the broken machine dataset, the weighted average of the F-score more than 5% higher. On the other datasets, the advantage in the proportion of correctly classified objects using RandomForest is not so big. It can be seen that the use of the mutual information criterion on the Email spam dataset degrades the quality of classification compared to classification without feature selection. variance.In general, the results with and without selection are practically the same, however, it can be noted that a consistently high-quality classification is also possible using the Chi-square and F-statistics analysis of

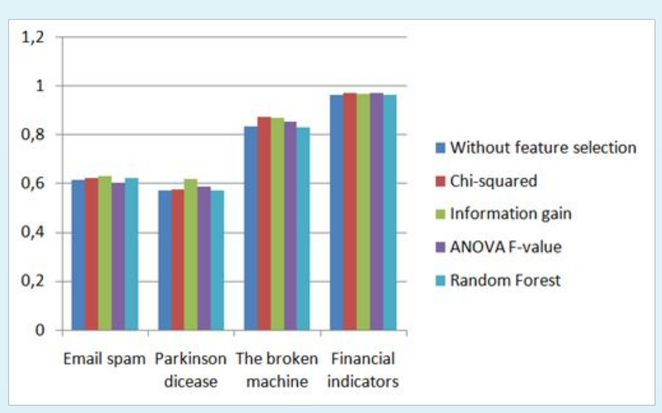

Fig. 2. Diagram of classification using logistic regression

TABLE II. Results of classification using Random Forest

	Email spam	Parkinson disease	The broken machine	Financial indicators
Without feature selection	0,61295	0,570025	0,830825	0,9614
Chi-squared	0,6217	0,5748	0,8704	0,9676
Information gain	0,6304	0,6163	0,8667	0,964
F-statistics ANOVA	0,6012	0,5877	0,8507	0,9675
Random Forest	0,6199	0,5698	0,83	0,9604

The diagram shows that the highest quality classification using a random forest occurs when features selected using the mutual information criterion, the proportion of correctly classified objects is approximately 2.5% higher than classification without feature selection is used. On the Parkinson Disease dataset, this statistic improves the classification result by 4.5% compared to classification without feature selection. The rest of the datasets also have some advance in classification. It is important to note that feature selection using RandomForest when using a random forest classifier shows a worse result compared to statistical criteria than it is used with the same classifier in combination with the SelectKBest algorithm. In general, the results with and without feature selection are practically the same, however, it can be noted that on all data sets a fairly high-quality classification occurs when using the Chi-square test and the mutual information criterion.

Conclusion

As a result of the work, various algorithms and criteria for feature selection were used to classify data. On average, feature selection gives a slight improvement in classification. It is important to note that using some criteria on some datasets may give a worse classification quality than before without feature selection. Based on the results from the tables, it can be concluded that feature selection using RandomForest is the most effective, the results of selection using this algorithm on almost all data sets and classifiers are better than without using feature selection. In addition, feature selection by this method gave the largest gain in the weighted average of the F-score compared to classification without feature selection.

On average, the proportion of correctly classified objects increased by 1% when using this algorithm. All statistical criteria in the best feature selection algorithm show the same results on average.

Thus, we can conclude that feature selection using a random forest in the general case, if a detailed analysis of the data set is not possible, will show the best result, however, it may turn out that the classification is more accurate on this data set when choosing the k best features using some statistical criterion, however, the choice of criterion requires a detailed study of the dataset but it is not always possible.

Acknowledgements

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

References

[1] J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang and H. Liu “Feature Selection: A Data Perspective” in ACM Computing Surveys, vol. 50, 2017, pp. 94-139

[2] Ходашинский, И.А. Отбор классифицирующих признаков: сравнительный анализ бинарных метаэвристик и популяционного алгоритма с адаптивной памятью / И.А. Ходашинский, К.С. Сарин // Программирование. – 2019. – Т. 45, № 5. – С. 3 - 9. DOI: 10.1134/S0132347419050030

[3] N. AlNuaimi, M. Masud, M. Serhani and N. Zaki “Streaming feature selection algorithms for big data: A survey” in Applied Computing and Informatics, vol. 18, 2019, pp. 113-135

[4] W. Mostert, K. Malan, and A. Engelbrecht “A Feature Selection Algorithm Performance Metric for Comparative Analysis” in Algorithms, vol. 14, 2021, pp. 100-116

[5] N. Hasan and Y. Bao “Comparing different feature selection algorithms for cardiovascular disease prediction” in Health and Technology, vol. 11, 2020, pp. 49-62

[6] Email spam classification dataset [Electronic resource]. — Access mode: https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv (05.02.2022)

[7] Parkinson disease speech signal features [Electronic resource]. — Access mode: https://www.kaggle.com/dipayanbiswas/parkinsons-disease-speech-signal-features (05.02.2022)

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

[9] 200 + financial indicators of US stocks (2014-2018) [Electronic resource]. — Access mode: https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018 (05.02.2022)

[10] K. Kirasich, T. Smith and B. Sadler “Random forest versus logistic regression: a large-scale benchmark experiment” in SMU Data Science Review, vol. 1, 2018

[11] M. Hasan, M. Nasser, S. Ahmad and I. Molla “Feature Selection for Intrusion Detection Using Random Forest” in Journal Of Information Security, vol. 7, 2016, pp. 129-140

[12] S. Vora and H. Yang “A Comprehensive Study of Eleven Feature Selection Algorithms and their Impact on Text Classification” 2017 Computing Conference. (18-20 July 2017, London, United Kingdom), 2017, pp. 440-449

[13] Y. Liu, J. Bi and Z. Fan, “Multi-class sentiment classifcation: The experimental comparisons of feature selection and machine learning algorithms” in Expert Systems with Applications, vol. 80., 2017, pp. 323-329