Comparison of Efficiency, Stability and Interpretability of Feature Selection Methods for Multiclassification Task on Medical Tabular Data

Ksenia Balabaeva[1] and Sergey Kovalchuk[1]

[1] ITMO University, Saint-Petersburg, Russia
kyubalabaeva@gmail.com
sergey.v.kovalchuk@gmail.com

Abstract.
Feature selection is an important step of machine learning pipeline. Certain models may select features intrinsically without human interactions or additional algorithms applied. Such algorithms usually belong to neural networks class. Others require help of a researcher or feature selection algorithms. However, it is hard to know beforehand which variables contain the most relevant information and which may cause difficulties for a model to learn the correct relations. In that respect, researchers have been developing feature selection algorithms. To understand what methods perform better on tabular medical data, we have conducted a set of experiments to measure accuracy, stability and compare interpretation capacities of different feature selection approaches. Moreover, we propose an application of Bayesian Inference to the task of feature selection that may provide more interpretable and robust solution. We believe that high stability and interpretability are as important as classification accuracy especially in predictive tasks in medicine.

Keywords: Feature Selection, Bayesian Inference, Explainable artificial intelligence, XAI, eXAI, recursive feature elimination, kbest.

1 Introduction

Due to the recent advances in machine learning algorithms application to different domains, there is a huge demand of decision support systems based on learning methods. One of the most popular type of machine learning tasks is supervised learning which undermines the use of a dataset with corresponded labels to each instance. Such tasks are, for example, regression, binary and multiclassification, depending on the target variable type.

In supervised learning, the input data typically consist of a matrix of features and a target vector. Such matrix may take a form of an image in the tasks of image classification, object detection or semantic segmentation. Another example of feature matrix may be vectorized text representation. Such feature matrix is widely

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

used in natural language processing tasks, such as, sentiment analysis, text classification, etc. Data in the form of images, video, plain text and audio is usually called unstructured data. However, the most widely used type of the input data is a tabular data containing variables of different nature in each column. Estimates say that 20% of the data are structured and approximately 80% are unstructured [1].

Machine learning algorithms' performance strongly depend on the number and variability of samples provided in training datatset. However, the raise in the number of variables may lead to the curse of dimensionality [2]. This problem refers to a higher risk of overfitting especially if the number of features is higher than the number of samples. Another problem arising from a big dimensionality is that the observations in high dimensional space become equidistant which makes them harder to cluster or classify. To solve this problem, we have to reduce the feature space. There are several ways to deal with it: feature extraction (PCA [3], LDA [4], Transformer [5]) or feature selection, which will be discussed in more details in further sections. All in all, the aim of all techniques is in reducing the number of columns in a training dataset [6].

In present study we analyze only feature selection selection algorithms and there are two reasons why we eliminate feature extraction methods. First of all, we would like to compare the existing feature selection methods with the proposed application of Bayesian Inference to this task. Since the proposed algorithm select features it is clear that at first step, we have to compare its performance with analogous methods. Another reason is that feature extraction approaches compress the initial feature space to reduce the dimensionality. For instance, using PCA we get a number of principal components in which initial features are encoded. After such compression we can't explain what features are contained in a single component. Even though the number of such components may be low, this way of compression causes difficulties for the interpretability. Compared to feature extraction, feature selection techniques are more transparent and explainable, since the reduced feature space consist of the original variables in data. Moreover, such reduction may reduce the training time and contribute to the accuracy of the model.

In the present work we compare several feature selection techniques on the case of chronic heart failure stage prediction. We also present an approach of Bayesian Feature Selection application to the task of feature selection. As comparative standards, we evaluate the selection algorithms using f-score with macro averaging as performance indicator, stability of the model, using k-fold cross-validation and interpretability of the feature selection results.

The rest of this paper is structured in the following way. Section 2 provides background on feature selection studies in medical domain, describes feature selection concepts and the most wide spread methods. Section 3 presents an approach to feature selection using Bayesian Inference. Section 4 provides details on experimental pipeline. Section 5 describes the results and discussion. Section 6 concludes the work.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

2 Related Works

The field of feature selection is a big part of machine learning domain. Therefore, there are plenty of algorithms appearing each year. However, there is lack of papers comparing the efficiency of feature selection techniques, their stability and interpretability. Therefore, it is quite hard for the practitioners to select the appropriate solution and understand the risk of overfitting using one or another method. Another issue is the lack of overviews on application feature selection methods to the medical domain.

As a rare example of such works, we may take the paper [6]. This work covers the field of medical imaging, DNA analysis, biomedical signals processing and testing the feature selection techniques on two tasks. As feature selection algorithms they compare CFS, Cons, INTERACT, InfoGain, ReliefF, SVM-RFE. There are also some works that address feature selection to one specific task. For instance, medical image retrieval [7] and In [8] the authors compare different information retrieval methods for medical image segmentation [8]. Such as thresholding based, clustering-based, watershed-based and graph-based, etc. In [9] authors apply Discrete Wavelet Transform to decompose an image into images with different scales in order to extract information. Another study concludes that biologically informed feature selection methods applied for Alzheimer diagnosis stages prediction are more efficient than uninformed [10]. Concerning medical signals processing, there are also a couple of works on EMG, EEG and ECG [11-12]. For instance, in [11] authors extract features from EMG signals using time, frequency and time-frequency domain features. Another domain of feature selection application to medicine is microarray data classification [13-15]. In [13] authors provide a review on feature selection methods, and include the software overview for microarray data, which is primarly written for R and C programming languages.

2.1 Feature Selection Approaches

Feature selection is a procedure of processing initial dataset in the end of which a sample of features become eliminated due to their redundancy and lack of useful information. There are several types of classification of feature selection approaches, but we decided to use the classification proposed in this work [16].

2.1.1 Filter Methods

Filter methods are techniques in which only features characteristics are used without a learning model. [17]. Generally, such approaches consist of two stages: choosing a criteria to rank the feature, and selecting top-ranking features. Such algorithms are Correlation-based Feature Selection (CFS) [18], Variance threshold [19], F-test [20], etc.

Variance Threshold removes features with variation below a certain threshold. Variance here is treated as an indicator of information provided by the feature: those features that do not change much across observations have less information. The

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

limitation of such method if that it doesn't undermine relations between features and labels.

Select K best. This approach is associated with statistical F-test, which conducts a hypothesis testing. As a limitation of this method, we have to admit, that it only checks for linear relationships between features and target variable. Another specificity is that features with high correlation will be given a higher score and less correlated features will get lower score.

General limitation of filter techniques is the inability to get the information from model's performance while selecting the optimal feature set.

2.1.2 Wrapper Methods

The class of the wrapper methods define the best feature subset as a subset that leads to the higher performance of the learning model. Therefore, such methods use specific learning algorithm to select features. [21]

. The work of a basic wrapper model could be divided into three stages: feature set selection, feature set evaluation and induction algorithm (predefined learning model).

Forward feature selection. On the first iteration the model with no features. Then the features are added to the training dataset one by one to reach the highest score [21].

Backward feature elimination. On the first iteration the model is trained on the whole dataset. Iteratively features are eliminated one by one in the way to get highest score [21].

Recursive feature elimination (RFE). This is an optimization algorithm which aims to find the best performing feature subset. Unlike previous methods, this approach creates new model recursively [22].

2.1.3 Embedded Methods

There is a class of feature selection methods that took the advantages from both filterbased and wrapper methods. Embedded models incorporate the feature selection process inside the learning model [23]. Embedded models are more robust than wrapper methods, since the selection process is not evaluated by the learning model. For that reason, they are more stable and less susceptible to over-fitting. The most popular embedded methods are ridge and lasso regression.

Lasso Regression. This method incorporates L1 regularization by adding penalty equal to the absolute value of the magnitude of regression coefficients

Ridge Regression. This method performs L2 regularization which by adding penalty equal to square of the magnitude of regression coefficients.

3 Methodology

As an alternative to existing feature selection algorithm, we propose to apply technique based on Bayesian Inference and probabilistic modeling. A similar procedure was applied to clustering results interpretation in our previous work [24]. However, the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

same technique can also contribute to feature selection in binary or multiclassification tasks.

The proposed approach is based on Bayesian inference [25]. The algorithm consists of three stages: posterior sampling, comparison matrix calculation and identification of features. The idea of this method is to select most typical features in each class by comparing their sampled distributions [24]. For instance, if the distribution of feature 1 significantly differs in class A compared to class B, we have to add feature 1 to the model. Based on this comparison we may select more relevant features that may help the classifier to build more accurate model.

Considering the methods classification provided in chapter 2 the proposed approach is associated with filter methods, because it doesn't rely on score of the learning model while selecting the features.

3.1 BI feature selection algorithm

Let $x_{n_i c_k}$ be a number of successes an observation belongs to class c_i with n_i being an overall number of observations for cluster c_i and p_{n_j, c_i} is the probability an observation belongs to class c_i.

Step 1. Posterior Sampling

For each variable f_j , $j \in [0, M]$:

For each class c_i , $i \in [0, K]$:

Select the priors for $x_{n_i c_l}$ and $P(A)_{f_j c_l}$;

Calculate the posterior distribution $P(A/D)_{f_j c_l}$;

Sample W new observations from the posterior $P(A/D)_{f_j c_l}$;

Step 2. Feature comparison matrix

Let I be a 2D matrix with the number of rows and columns equal to the number of classes. The value of each matrix element $i_{c_l c_{l+1}}$ is equal to the mean value of sampled probabilities comparison. The calculation depends on the hypothesis we want to check: whether the values of a feature in one class are higher or lower to the features in other class.

Step 3. Identification of features more typical for a class

Output: dictionary with keys equal to class numbers and values equal to array of associated features with this class [24]. Finally, we can build a sample of more relevant features concatenating the output of the third step.

This approach has two main parameters: the significance level and the number of classes to which the distribution of a feature in current class must be significantly different. The significance level may vary from 0 to 1, where 0 means that there is no differences between distributions in classes and 1 means that the distributions are completely different. The number of comparison classes may vary from 0 to the maximal number of classes in target vector minus 0.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

4 Experiments. Case of Chronic Heart Failure Stage Prediction

To test feature selection methods, we picked a multiclassification task, training ML models to predict the stage of congestive heart failure. Congestive heart failure (CHF) occurs when heart muscle struggles with pumping the blood as well as it should. This disease may be caused by narrowed arteries in heart or arterial hypertension which is a widely spread chronic disease. According to the clinical classification, there are 4 stages of CHF, where the first stage represents the weak disease and the fourth represents the severe progression of CHF.

The dataset consists of 1279 observations represented by patients. The target vector has 4 classes, representing the stages of CHF. Distribution of the target vector is depicted on figure 1. The most popular stage in the sample is the third one – there are more than 600 patients with this stage.

Figure 1. Distribution of CHF stages

The feature dimension is represented by socio-demographic characteristics (age, gender), labs results (hemoglobin, neutrogene, etc.), blood pressure measures, main diagnosis, etc. In total, there are 178 features describing each patient, which is an extremely high dimensional space.

In order to reduce the number of features, we test different feature selection techniques (Table 1). Each of the selection algorithm has its own parameters that we optimized according to our quality metric F1_score.

Table 1 Feature Selection Algorithms and their parameters

Feature Selection Technique	Parameters
Variance Threshold	Threshold
Bayesian Feature Selection	Significance Level, Num Classes
KBest	Number of Features Selected

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

RFE	Minimal Number of Features to select
Lasso Regression	-
Ridge Regression	-

Since Ridge and Lasso Logistic Regressions are embedded methods, we used them only within Logistic regression, treating each modification as a single classifier.

After the selection is completed, we pass the new feature set to ML classifier. For this task three ML models were tested: Logistic Regression (Ridge and Lasso), Random Forest and Gradient Boosting. Further we check the quality of predictions using the test dataset and cross-validation with 5 folds calculating f1-score with macro-averaging. We chosen f1-score because it can be used for imbalanced classes (figure 1). The results of the experiments are presented in chapter 5.

5 Results and Discussion

According to the experimental pipeline, we compared classifiers performance applying different feature selection techniques (Table 2). The scores presented in table 2 were calculated on test data (33% of the initial dataset). The highest f1-score on test set for all classifiers was performed by recursive feature elimination (RFE). The second-best feature selectors were Bayesian Selection for Logistic Regressions and KBest (F-test) for Random forest and XGBoost.

Table 2. Comparison of the feature selection techniques efficiency according to the model's f1 score based on test data

Classifier/Selection Technique	NO Feature Selection	Variance Threshold	Bayesian Selection	KBest	RFE
LogReg	0.36527	0.36527	**0.39669**	0.36527	**0.48699**
LogReg + L1	0.3793	0.3793	**0.39669**	0.3793	**0.4869**
LogReg + L2	0.36527	0.36527	**0.39669**	0.36527	**0.48699**
Random Forest	0.45875	0.37113	0.4449	**0.46299**	**0.5839**
XGBoost	0.46013	0.39095	0.45019	**0.47764**	**0.63593**

We used K-fold validation to check stability and robustness of the classifiers trained on the selected feature sets. However, it is a useful tool to measure accuracy as well. The validation results are presented in table 3. Here we see that almost all of the selection methods are losing the quality being checked on validation samples. However,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

the most significant drop is associated with RFE (0.15-0.20 f-1score decrease). According to the validation, the most accurate classification was performed using Bayesian Selection for logistic regression, K-Best selection for Random Forest and RFE for gradient boosting.

Concerning the stability, RFE is more exposed to the overfitting, since it is a wrapper method, and it exploits learning algorithms for feature evaluation. This issue finds confirmation in our experimental results due to the high score differences on training and validation sets. Other algorithms have relatively similar change in the score.

Table 3. Comparison of the feature selection techniques stability according to the model's mean validation f1-score (+- std).

Classifier/Selection Technique	No Feature Selection	Variance Threshold	Bayesian Selection	KBest	RFE
LogReg	0.34528 (+- 0.0442)	0.34528 (+- 0.02214)	**0.3565**(+- 0.0369)	0.34528 (+- 0.0442)	0.32973 (+- 0.0313)
LogReg + L1	0.34677(+- 0.0221)	0.34677(+- 0.04252)	**0.3565**(+- 0.0369)	0.34677(+- 0.0442)	0.34677(+- 0.03132)
LogReg + L2	0.34528 (+- 0.02214)	0.34528 (+- 0.02214)	**0.35652** (+- 0.0369)	0.34528 (+- 0.0442)	0.32973 (+- 0.03132)
Random Forest	0.38410 (+- 0.04653)	0.39642 (+-0.1555)	0.404787 (+- 0.08486)	**0.4105** (+- 0.14035)	0.39829 (+- 0.0329)
XGBoost	0.4220 (+- 0.0270)	0.370087 (+- 0.044835)	0.41719(+- 0.0702)	0.3022 (+- 0.06949)	**0.43086** (+- 0.0632)

The third criteria of comparison id the optimal number of features found by each feature selection algorithm (Table 4). To a certain extent, the number of features represent the complexity, transparency and interpretability of the solution. The smaller the number of features – the easier it is to explain the results. For all classifiers Variance Threshold selected 27 features with threshold equal to 0.9-0.7. That is the smallest number of features selected by any algorithm. The size of the feature sample selected by Bayesian Selection varies from 47 to 73 for different models. KBest algorithm picked from 50 to 140 features. And RFE selected 27 features for Logistic Regression Model, 44 for XGBoost and 90 for Random Forest. Even though for Logistic Regression both RFE and Variance Threshold picked the same number of features, the sample selected by Variance Threshold is more relevant, according to validation results (Table 3).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

Table 4 Comparing the number of features selected by each FS technique

Classifier/Selection Technique	No Feature Selection	Variance Threshold	Bayesian Selection	KBest	RFE
LogReg	178	27	73	140	27
LogReg + L1	58	27	73	140	27
LogReg + L2	178	27	73	140	27
RandomForest	178	27	47	50	90
XGBoost	178	27	64	90	44

Concerning the interpretability of feature selection techniques, filter based methods and embedded methods are more transparent to users, since the logic of feature selection is simpler. For instance, variance threshold is just the elimination of uninformative features. Or Bayesian Selection is just the selection of features more specific for each particular class. Whereas RFE is a complex process of feature selection and evaluation requiring learning models. We do not include the nomenclature of selected features, since their initial number is high – almost 200 features and the limits of the paper won't allow us to discuss each selected set in detail.

6 Conclusion

As the main result, we have to say that the proposed objective of our work is reached. We suggested a filter-based approach to feature selection task based on Bayesian inference and probabilistic modeling. This method performs sufficient accuracy for multiclassification task and is quite robust to the problem of overfitting. In our experiments we compared this method with other feature selection algorithms and presented the results concerning their stability, accuracy and explainability.

All things considered, the obvious conclusion to be drawn is that feature selection may help to improve the ML models performance, reduce the learning time and foster the search of the optimal hyperparameters due to dimensionality reduction.

However, the choice of the feature selection algorithm strongly depends on data and machine learning model. In our case, Bayesian feature selection performed better on Logistic regression, K-best algorithm reached higher score working with Random forest and RFE booster the performance of XGBoost.

In future we would like to compare the proposed approach with other types of feature selection methods and test it on different datasets.

Acknowledgement

This research was supported by the by the Ministry of Science and Higher Education of Russian Federation, goszadanie no. 2019-1339. Participation in the ICCS conference

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

was supported by the NWO Science Diplomacy Fund project # 483.20.038 "Russian-Dutch Collaboration in Computational Science"

References

1. https://www.cio.com/article/3406806/ai-unleashes-the-power-of-unstructured-data.html, last accessed 2021/02/10
2. Bellman R.: Dynamic Programming. Princeton University Press (1957)
3. Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh.. Principal Component Analysis. International Journal of Livestock Research. (2017) 1. 10.5455/ijlr.20170415115235.
4. Blei D., Ng A., Jordan M.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022
5. Polosukhin, I., Lukasz K., et. Al:. Attention Is All You Need. (2017).
6. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. (2019) Sep;112:103375. doi: 10.1016/j.compbiomed.2019.103375. Epub 2019 Jul 31. PMID: 31382212.
7. J. Kalpathy-Cramer,: Evaluating performance of biomedical image retrieval systemsan overview of the medical image retrieval task at imageclef 2004–2013, Comput. Med. Imag. Graph. 39 (2015) 55–61.
8. Q. Huang, Y. Luo, Q. Zhang, Breast ultrasound image segmentation: a survey, Int. J. Comput. Assist. Radiol. Surg. 12 (3) (2017) 493–507.
9. V.K. Sudarshan, M.R.K. Mookiah, U.R. Acharya, V. Chandran, F. Molinari, H. Fujita,K.H. Ng, Application of wavelet techniques for cancer diagnosis using ultrasound images: a review, Comput. Biol. Med. 69 (2016) 97–111.
10. S. Rathore, M. Habes, M.A. Iftikhar, A. Shacklett, C. Davatzikos, A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages, Neuroimage 155 (2017) 530–548.
11. N. Nazmi, M. Abdul Rahman, S.-I. Yamamoto, S. Ahmad, H. Zamzuri, S. Mazlan, A review of classification techniques of emg signals during isotonic and isometric contractions, Sensors 16 (8) (2016) 1304.
12. U.R. Acharya, H. Fujita, V.K. Sudarshan, S. Bhat, J.E. Koh, Application of entropies for automated diagnosis of epilepsy using EEG signals: a review, Knowl. Based Syst. 88 (2015) 85–96
13. Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
14. B. Remeseiro and V. Bolon-Canedo Computers in Biology and Medicine 112 (2019) 103375 8 bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517
15. V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J.M. Benítez, F. Herrera,A review of microarray datasets and applied feature selection methods, Inf. Sci. 28 2 (2014) 111–135.
16. Colaco S.: Review on Feature Selection Algorithms, (2019)
17. Liu, H., Motoda, H.:. Computational methods of feature selection. Chapman and Hall/CRC Press (2007)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51

18. Doshi, M., & Chaturvedi, D. S. K. (2014). Correlation based feature selection (cfs) technique to predict student perfromance. International Journal of Computer Networks & Communications (IJCNC), 6(3)
19. Guyon I.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 (2003) 1157-1182
20. Elssied, Nadir & Ibrahim, Assoc Prof. Dr. Othman & Hamza Osman, Ahmed. A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. Research Journal of Applied Sciences, Engineering and Technology. (2014). 7. 625-638. 10.19026/rjaset.7.299.
21. Kohavi, R., John, G. H.: Wrappers for feature subset selection. Artificial intelligence, 97(1), 273–324. (1997).
22. Sam, M. L., Camara, F., Ndiaye, S., Slimani, Y., & Esseghir, M. A. (2012 June). A Novel RFESVM-based Feature Selection Approach for Classification. International Journal of Advanced Science and Technology, 43
23. Kabir, M. M., Islam, M. M., & Murase, K. (2011). A new local search based hybrid genetic algorithm for feature selection. Neurocomputing, 74, 2194–2928
24. Balabaeva K, Kovalchuk S. Post-hoc Interpretation of Clinical Pathways Clustering using Bayesian Inference. Procedia Computer Science 178: 264-273 (2020)
25. Cameron, Davidson, Pilon. Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference. Addison-Wesley (2019)
26. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
27. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).
28. Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010).
29. LNCS Homepage, http://www.springer.com/lncs, last accessed 2016/11/21.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_51