DOI:10.34918/85112

Katarzyna SIERADZKA[1,*], Joanna POLAŃSKA[1]

Chapter 12. FEATURE SELECTION METHODS FOR CLASSIFICATION PURPOSES

12.1. Introduction

Nowadays, the development of the needs to generate larger and larger data sets is undoubtedly related to the development of technology and equipment that generates this data. It would seem that the more data we collect in a research problem, the more precise the description of the studied phenomena becomes. However, it often turns out that instead of generating more and more accurate and useful information in the analysis of high-dimensional data, we receive redundant and highly distorted information. Without a careful selection of these data, we expose our research to unwanted use of more and more computational resources and the time necessary to analyze even unnecessary and without practical information features. The problem of large amounts of data for analysis occurs with automated text and language analysis. As described by Justin Grimmer et al. [1] in the analysis of a large number of texts, it is not possible to manually read all the articles related to the problem. It is connected not only with a physical lack of time, but also with limited finances. Therefore, there is a need for automatic text analysis, with the use of complex algorithms dedicated to this problem. This phenomenon of data redundancy very often occurs also in the problem of generating data from single-cell sequencing experiments, where as a rule we get a lot of complex information about gene expression. This data is made up of tens of thousands of cells and hundreds or even thousands of features. The need for in-depth analysis and the use of automated methods to search for patterns relevant to the research problem for these data is of key scientific importance [2]. The use of decision support algorithms is therefore of great importance in healthcare, where artificial

[1] Department of Engineering and Data Exploratory Analysis, Silesian University of Technology, Gliwice, Poland.
[*] Corresponding author: katarzyna.sieradzka@polsl.pl.

intelligence analyzing real data is able to significantly assist healthcare workers in making decisions, and what is more, such systems can also warn about detected irregularities [3].

However, cooperation with self-learning artificial intelligence algorithms requires careful consideration of the analyzed features. The lack of control over these factors may lead to a significant extension of the waiting time for learning the predictive model, disturbances in the learning process and, consequently, to making wrong decisions [2]. The natural process of generating high-dimensional data is the formation of irrelevant and redundant features. Keeping control over them is therefore particularly important from the point of view of the analysis and computational costs. Due to the high significance and the need to eliminate features, undoubtedly adversely affecting the processes of machine learning and attempts to solve the research problem. In 1997 the first studies on feature selection (FS) were described, based on large data sets [4–5]. Since then, many FS techniques have been developed, and their development continues to this day. Particular workload is focused on the field of machine learning, where the appropriate selection and analysis of features describing the domain, classifying and recognizing significant patterns is the key to obtaining practically useful and interpretable results. The development of these techniques is of particular importance nowadays, when analyzing often tens of thousands of features at the same time [6]. Removal of worthless features enables a significant reduction in the costs and computational resources necessary to conduct an appropriate study. An alternative solution is also the introduction of ensemble feature selection (EFS) algorithms, i.e. a combination of several single FS techniques, which enables the integration of the advantages of many FS techniques while eliminating their disadvantages [6].

12.2. Overviev of irrevelant and redundant features removal techniques

A very frequently used technique in the initial stages of high-dimensional data analysis is dimensional reduction. Dimensionality reduction refers to the reduction of the number of features in a data set while maintaining information relevant from the point of view of the research problem. This technique can be divided into feature extraction and feature selection. The concept of feature extraction refers to the generation of completely new features, which are a combination of original features present in the input data set. This allows you to limit the dimensionality of the data by

selecting only those newly created features that will explain the problem to the greatest extent. However, it is very difficult to associate the newly created features with the input features later. The new features are in no way physically interpretable, as was the case with the original features. Feature extraction techniques are very well developed in a machine learning environment and include Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Canonical Correlation Analysis (CCA), among others [7]. On the other hand, feature selection techniques are based on selecting an appropriate number of features without transforming them. The problematic aspect in the case of FS techniques is the selection of this appropriate number of features. With the use of new methods of generating a very large amount of data, we obtain much more information that is important to us, but it should also be remembered that redundant and irrelevant features also arise to a much greater extent. The selected features should constitute a proportionally small part of the set of input features, and at the same time retain as much information relevant as possible from the point of view of the research problem. Feature selection as compared to feature extraction maintains the readability and interpretability of the data, because no transformations combining many features into one are used here. This is an undoubted advantage of the FS techniques, because the obtained sets of selected features have a physically interpretable meaning, which is particularly important when looking for, for example, expression patterns or genes influencing specific diseases [7]. Therefore, further considerations in this manuscript regarding dimensional reduction techniques will focus on feature selection techniques. Among them, we can distinguish three main types: wrappers, filters, and embedded techniques.

12.2.1. Wrappers

Wrapper is a method of selecting the optimal set of features by a learning algorithm. The choice of the classifier is arbitrary, but it is a method that requires a large amount of computational expenditure, and thus a large amount of time, due to the need to run the classifier repeatedly based on different feature subsets [6]. Due to the fact that computing the objective function is a computationally demanding task, wrappers are not a perfect solution for data with a high complexity of features [2]. In a very general approach, this method consists in using the quality of the prediction provided for the selected classifier to determine the usefulness of the analyzed subsets of features. The applied feature search strategies can be divided into a range of searching strategies e.g. hill-climbing , best-first, branch-and-bound, or genetic algorithms [8]. However, the most promising in terms of robust against overfitting, and at the same time the

most computationally complex, are greedy search strategies [7]. They are divided into two techniques of features selection: forward selection and backward elimination. Forward selection is the gradual addition of features to a subset of features that ensure the achievement of better and better classification results. In this type of feature selection, the subset of features begins at the empty subset of features and grows with each successive step, until the point where adding more features does not significantly affect the classification process. The opposite effect is seen in the case of backward elimination. In this case, in the following steps, from the full set of features, those with the least promising effect on the quality of the classifier are removed from the feature subset. In the case of the two methods described above, i.e. forward selection and backward elimination, one can often find different opinions about which method gives the best results. Which method we use can of course be defined by the purpose of the research being carried out. The use of forward selection allows us to choose from the set of features the one feature that best allows the separation of variables – the forward selection selects the best differentiating feature in the first step, adding the next one in the next step, which in cooperation with the first one gives better results. On the other hand, backward elimination, due to the fact that it starts on the full set of features, rejects the one feature that contributes the least to the interaction of the remaining features in the set [8]. The earlier dimensionality reduction could seem to be a solution that allows to significantly reduce the time of FS procedures using wrappers methods. However, the selection of a significant, in terms of a complex and difficult research problem, subset of features on a previously limited subset may adversely affect the results of this method. The features that were previously eliminated could potentially be included in a subset selected with the use of wrappers methods, significantly improving the quality parameters of the classifier. In this case, the dimensionality of the data can be reduced using simple linear data transforms such as PCA or LDA, as well as more sophisticated ones, such as the Fourier transform, while (as mentioned in the introduction) it deprives the possibility of a physical interpretation of a selected subset of features [8].

The problem of selecting features is not only based on the selection of the smallest possible subset of features that will enable the achievement of average classification results. Particularly in areas where machine learning is intertwined and cooperates with genetics or health care, one of the most important aspects is data interpretability, finding disease patterns, connections of individual features, signal pathways, and recreating a certain biological history that stands behind the actual state of health. However, in the case of data with very large dimensions, and we have to deal with such data very often, the feature selection path should be very seriously considered.

As Isabelle Guyon et al. writes [8], this path is based on using a linear predictor of choice, and consists in the first step of preselecting features using simple statistical methods such as correlation and ranking of features, and in the second step using forward selection or backward elimination methods. Choosing such a path with several thousand features may significantly accelerate the computational processes, and does not have to deteriorate the results for the selected subset of features in a statistically significant way.

12.2.2. Filters

The second FS method are filters. Filters as a method of FS is not based on classification, as was the case with wrappers [9]. Therefore, they allow to perform the necessary, to select a subset of the features, analysis in a shorter time, so they can be utilized in the case of high-dimensionality datasets [6]. These methods are quite simple in terms of computational complexity and generally consist of two basic steps. In the first step, features are ranked using an appropriate criterion. This step can be performed both on individual features and on sets of features. In the next step, features with the highest ranks are selected for a subset of features, on the basis of which further analysis and inference are possible [7]. Many algorithms have been developed that play an important role in selecting filters as FS methods. Moreover, these methods can be divided into unsupervised and supervised [2]. Among the unsupervised methods, we can distinguish: term-variance (TV) criterion [10], Laplacian score (LS) [11], and Fisher score. TV score sorts the features according to their variance in the given samples. LS use the local geometric structure to assess the significance of a feature [2]. In the Fisher score, individual features are treated separately, so it is not able to detect the relationship between the individual features. Thus, it is unable to detect redundant features [7]. In order to solve this problem, Gu et al. [12] proposed a generalized Fisher score in which features are selected together and it is possible to maximize the lower bound of traditional Fisher score. Another approach to using filters in FS methods is the mutual based information method. This method is based on counting the information gain between successive features and labels. The feature is significant if it has a high information gain. In this method, as in the traditional Fisher score approach, features are selected in univariate way and it is not possible to determine redundant features. The second group of filters are supervised methods such as Relief [13]. It is an algorithm based on the measure of significance of features, the value of which is related to how well the values of this

feature distinguish instances of the same and other classes. This method is very useful in the analysis of data of high complexity, but it does not allow the removal of redundant features [2]. In the ReliefF method, extended by the multiclass problem, compared to the Relief, which was used to handle two-class problems, the features are selected in such a way as to separate the individual randomly selected, in the subsequent stages of the algorithm's operation, elements of the data set coming from different classes [13]. In this case, the quality of the features is calculated on the basis of how well their values are able to distinguish between individual instances located close to each other [14]. Another, also well known, FS algorithm is the fast correlation-based filter (FCBF) [15]. This method is based on the significance of features, but cannot be successfully applied to high-dimensional data [2]. For this reason, a new solution was introduced, such as the minimum redundancy maximum relevancy (mrMR) criterion [16]. It focuses both on redundancy between features and the importance of each feature. As Artur J. Ferreira et al. [2] writes, many algorithms do not work well with high-dimensional data. They proposed a new filter-based procedure dedicated to the large-dimension data approach. The first proposed algorithm uses the dispersion measure to determine the relevance of each feature, and then sorts them in descending order to preserve the selected number of features with the highest value of the measure used. The second algorithm, however, when more than one feature turns out to be redundant, is responsible for selecting only one of them, making comparisons of subsequent features. These comparisons, for the sake of saving computing time, are performed only for the top relevant features. As the authors write, this method removes the most redundant features from a subset of the most relevant features. What is worth emphasizing, both algorithms can work in unsupervised or supervised mode.

12.2.3. Embedded methods

Embedded methods are another method used in FS procedures. Their main goal is to derive the best results from the learning process based on a subset of features. For this reason, these are methods that combine the previously described wrappers and filters [6]. The feature selection algorithm is integrated as part of the learning algorithm. A learning algorithm takes advantage of its own features selection process and performs FS and classification at the same time. Embedded methods combine the advantages of wrappers, i.e. they take into account interactions with the selected

classification and filter model, i.e. they are much less computationally demanding than wrappers [7, 17–19]. Currently, there are many variants of embedded methods for the FS problem. These are, for example, C4.5 programs [20] to create decision trees, random forest [21], as well as algorithms based on multinomial logistic regression with its variants [22]. The random forest methodology is based on the combination of many decision trees. Each individual tree is created on the basis of a random sample selected from the input data set. For each tree, in each node, successive divisions occur on the basis of randomly selected traits, which are used as candidate traits to make the best division. In the course of subsequent divisions, the statistics of the significance of features are calculated, among others [23]. Other methods are based on regularization models which by minimizing the matching errors in a short time make it possible to set some feature coefficients to very small or even exact zero values [24, 25]. Methods that penalize features that do not affect the model's performance are typically those that work with linear classifiers (for example, SVM [26]), such as LASSO [24]. LASSO is the Least Absolute Shrinkage and Selection Operator method utilized in linear models, but can be applied in another statistical models including tree-based and generalized regression models [27]. LASSO uses a regularization procedure to reduce the value of the regression coefficients. In the process of selecting features after regularization, variables with non-zero values are selected for the model. In this case, the lambda parameter is of great importance, as it is used to control the strength of penalties, and thus affects the number of zero regression coefficients [28]. Despite the many advantages of embedded methods, it should be noted that we are not able to calculate the significance value of features for all types of machine learning algorithms, for example for nearest neighbors method [19].

12.3. Classification problem

The selection of traits mainly affects the training phase of the classification process. The FS process can be completely independent of the learning process (filters), but it can also be built into an algorithm that iteratively evaluates the significance of individual selected features (wrappers) [7]. The feature selection is of key importance, because its task and assumption is to select a subset of features that in the best possible way discriminates against observations belonging to different classes. Thus, the significance of a given trait can be described by its ability to distinguish between

classes. When trying to define the concept of classification itself, it can be concluded that it is a problem of assigning an unknown observation to a specific class, based on the training set for which the belonging of observations to particular classes is known. In the learning process, the algorithm uses information about selected features and belonging of individual observations to specific classes to build the function of assigning observations to classes. Having this function, in the prediction process, the classifier assigns new observations to membership classes based on the previously collected information. A very important element related to the classification problem is the validation set, which should not participate in any of the stages of teaching the classifier. This set should be randomly selected before the information is introduced into the learning process. This will ensure the independence of this set and enable a real assessment of the quality of the final classifier performance. An additional advancement in the learning process is the drawing from the rest of the training set dataset and repeating the training procedure on randomly selected subsets. This procedure can be essential for very small sample sizes. The assessment of what proportion of the input data should be the test set and which part of the training set is unclear. In the case of small sets, it is possible to apply leave-one-our cross-validation, which consists in drawing one observation from the set and then testing it. The rest of the set is for training. This method, due to a very large generalization in the selected observation, is burdened with a very large evaluation error [8, 29]. In the case of classifiers, we can of course use both supervised and unsupervised methods. The supervised selection of features determines their significance based on information about the class affiliation of observations, but for the learning process to be successful, there is often a need to provide a large number of observations with assigned affiliation labels, which is time-consuming. On the other hand, unsupervised selection works, of course, based on observations without assigned labels, but in this case the assessment of the significance of the features is not straightforward and is difficult [7].

In general, there are four basic and necessary steps to consider [17]: feature selection, feature evaluation, stop criterion, and validation. In the first step, using an appropriate method, a subset of features is selected, which in the next step is assessed according to a specific criterion. The selection of significant features is completed when the assumed stop criterion is reached. The selected subset is evaluated against the validation set.

12.4. Proposed method

The process of selecting a subset of features, that will be utilized to build a final model for heterogeneous data set, proposed in this manuscript is a two-step process. The first step is to use the wrapper method, more specifically the forward selection.

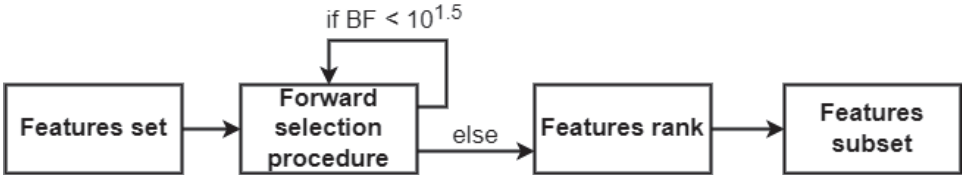

Fig. 1. General outline of presented method for features selection problem
Rys. 1. Ogólny zarys zaprezentowanej metody selekcji cech

The second stage is the use of the filter methodology and the introduction of the feature ranking measure on the subset previously pre-selected by the forward selection algorithm. A general diagram showing the individual steps of this method is presented in Fig. 1. In the forward selection procedure, Bayes Factor (BF) was used as a measure informing about the profitability of keeping the model with a higher degree of complexity (more features included in the proposed model). The BF value was calculated from (1).

$$BF = e^{(LogLikelihood_i - LogLikelihood_{i-1})} \tag{1}$$

where:
i means the level of models' complication.

In the initial set subjected to the FS procedure, there were 406 features. The initial data consisted of 4,214 items, among which two classes were distinguished. In the control class, there were 2,252 elements, and in the class subjected to the research procedure, hereinafter referred to as the procedural class, 1,962 elements. A very important step, which was carried out in the initial phase of the analysis, was to determine the elements representing the validation set. These elements were randomly selected from both classes, maintaining the dependence of 20% of the number of elements in these classes. The remaining part of the input data created a set which was subjected to the procedure presented in Fig 1. The numbers of elements broken down into classes and generated sets are presented in Table 1.

Table 1

Number of elements in both randomly generated sets

	FS procedure	Validation
Control class	1800	452
Procedural class	1571	391
Overall	3371	843

12.4.1. Forward selection procedure

The selection of a subset of features consists in running the model building, logistic regression (LR) methods based, algorithm 50 times. In the initial phase of the procedure, the training and test set are drawn. Importantly, the test set is balanced, each time the algorithm is run, in terms of the number of elements from the control and procedural class. 1012 elements were randomly selected for the test set (506 elements from each of the two classes). The remaining part of the set subjected to the feature selection, i.e. 2359 elements (1294 from the control class and 1065 from the procedural class), was the training set. This set was then subjected to the feature selection stage using forward selection methods. In the first stage of model building, single-element models were created taking into account each of the available features. The values of the N parameters of the generated models were calculated until a sufficiently small difference was obtained between successive likelihood value estimates for the model with a given set of parameters. All N univariate models were then compared using the Bayesian Information Criterion (BIC) measure described by (2). The best model, with the lowest BIC value, was recorded along with the estimated log-likelihood value. In the next step, we started building models composed of two features.

$$BIC = N_{parameters} \times \ln(N_{cells}) - 2 \times \mathrm{LL} \quad (2)$$

where:

LL is the log-likelihood function.

For this purpose, each of the remaining N-1 features was attached to the feature selected in the previous step, which resulted in the generation of N-1 two-element models. The procedure for selecting the best two-factor model remained the same as for the one-factor models. The BIC values of each of the generated models were compared and the best one was selected. The BF measure was used to determine whether the previously selected one-factor model or the two-factor model calculated in the current iteration will be kept. A threshold indicating strong evidence in favor of

a more complex model was selected as the value determining the next stages of the procedure (the exact value of the BF threshold is shown in Fig. 2). If a two-factor model is selected, the three-factor model building procedure is started and the procedure is the same as that described above. If a model with a less complexity is selected, the procedure of adding further features to the model is completed. In a critical case, the procedure can also be completed when all available features in the input set are used.

After receiving 50 models generated in this way, testing is carried out with the use of randomly selected, in subsequent runs of the algorithm, balanced test sets. The entirety of the described procedures leading from the input set of features to model testing is presented in Fig. 2.

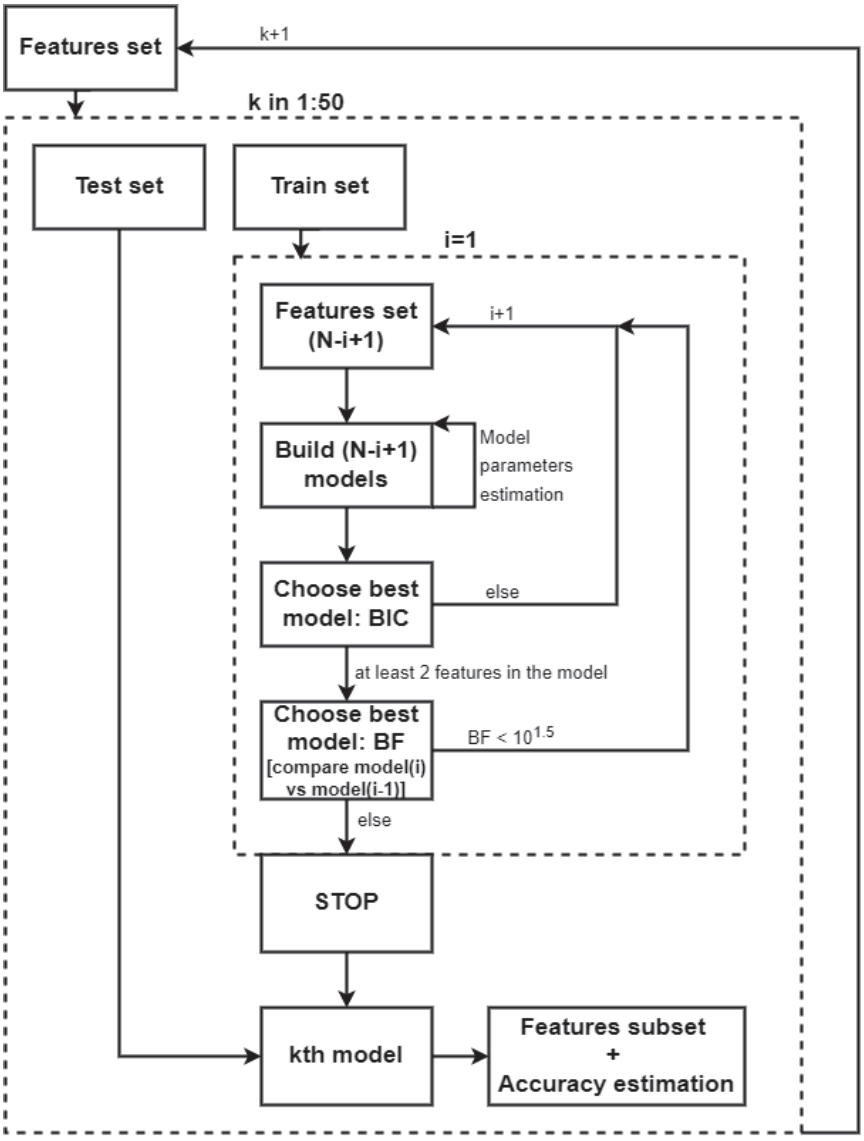

Fig. 2. Forward feature selection procedure. Where: *k* is the algorithm run ID, *N* is the total number of features, and *i* is the number of genes in model

Rys. 2. Procedura selekcji cech w przód. Gdzie: k to identyfikator uruchomienia algorytmu, N to całkowita liczba cech, i to liczba genów w modelu

12.4.2. Features rank

In the second step of FS, the filters method was used, based on the ranking of features. The ranking was created on the basis of 50 models generated with the use of logistic regression methods. The FeatureRank measure from (3) is based on 3 key elements related to individual features across all generated models: classification quality for a given feature related to the test set, number of features in the longest model, and the position of a feature in the model. Each feature that appeared at least once in one of the 50 models was taken into account in the ranking. The enumerated FeatureRank values were then normalized to a range of 0 to 1.

$$FeatureRank = \sum_{j=1}^{N} \frac{accuracy_{test_j} \times (k - i_j + 1)}{k} \tag{3}$$

where:
N is the models' ID, k is the number of features in the longest model, i is specified features' position in the model, and $accuracy_{test}$ is the estimated accuracy value for testing set for specified model.

With the use of the ranking of features, there has also been a need to determine the appropriate number of features that are important. For this purpose, the cut-off threshold was set before the features that showed a decrease in the differences in the estimated FeatureRank values. In other words, the features were rejected which not only had very low values of the measure assigned, but also the differences between these values for the following features were insignificant.

12.5. Feature selection results

Table 2

Estimated accuracy values based on testing set

Model ID	Testing accuracy [%]	Number of features	Model ID	Testing accuracy [%]	Number of features
1	91.25	31	**26**	91.87	31
2	92.27	33	**27**	92.20	31
3	90.32	27	**28**	89.57	32
4	92.99	30	**29**	90.81	37
5	90.35	26	**30**	90.43	33

cont. table 2

Model ID	Testing accuracy [%]	Number of features	Model ID	Testing accuracy [%]	Number of features
6	91.31	34	**31**	90.63	22
7	90.24	31	**32**	91.65	29
8	91.51	30	**33**	**89.37**	25
9	92.61	28	**34**	91.22	23
10	**93.59**	37	**35**	89.95	21
11	90.13	21	**36**	89.88	34
12	91.32	28	**37**	91.95	28
13	91.09	24	**38**	90.43	31
14	91.99	30	**39**	91.38	24
15	91.24	23	**40**	91.22	28
16	90.52	29	**41**	91.60	36
17	91.54	30	**42**	90.91	31
18	91.61	29	**43**	91.12	36
19	92.31	39	**44**	91.40	22
20	93.58	32	**45**	93.43	33
21	91.57	39	**46**	91.13	28
22	89.97	29	**47**	92.69	28
23	92.43	36	**48**	90.82	33
24	91.04	26	**49**	91.91	35
25	91.91	27	**50**	90.24	29

After using the logistic regression-based classifier on the input set of 406 features the average value of the classification quality, for the test sets for 50 models, was obtained at the level of 91.33% with the 95% confidence interval (91.05÷91.61). The minimum value for quality is 89.37% and the maximum is 93.59%. Table 2 describes the obtained classification qualities for the test sets for all the created models. Importantly, in this case, the quality of the classification was not counted as the weighted quality of the classification due to the fact that the test set was balanced in each subsequent draw in terms of the number of elements from the control class and the procedural class. Additionally, there was also estimated the 95% confidence interval for the mean number of features included in specified models which equals (28.51÷31.05).

After receiving a set of 50 models with information on selected features, the FeatureRank values were calculated for each feature that appeared in any of the models at least once. Calculations of this measure were performed for all models simultaneously. 159 features entered this stage of feature selection, based on filter

methods. Each of them was assigned a FeatureRank value ranging from 0 to 1, and then the features were ordered in descending order of significance measure. The ranking of the features is presented in Fig. 3 below.

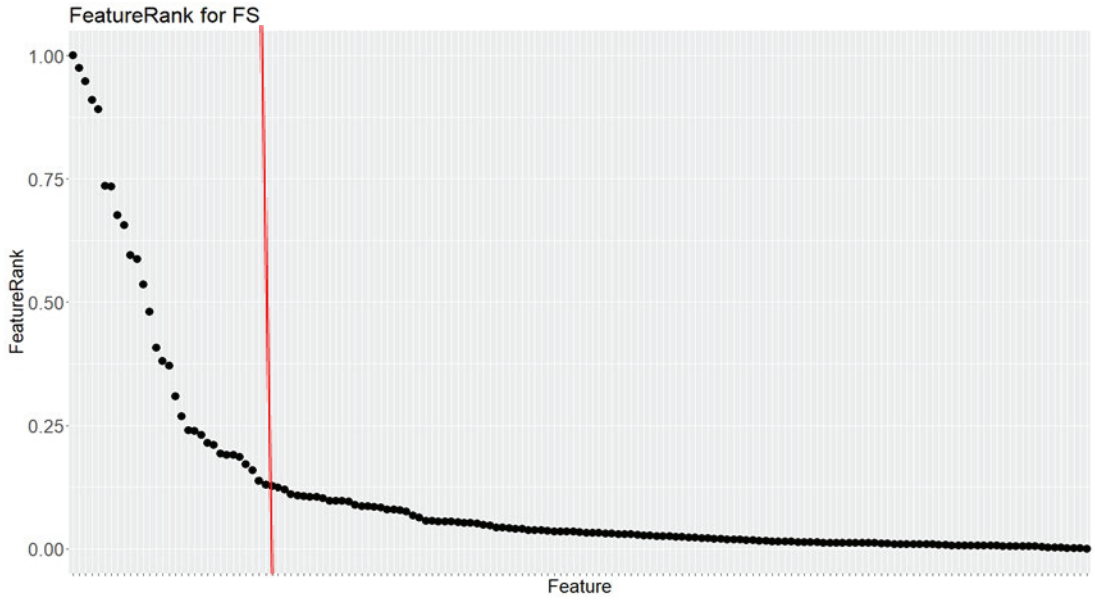

Fig. 3. Ranked FeatureRank metric with marked threshold for the number of selected features
Rys. 3. Miara FeatureRank z oznaczonym progiem odcięcia dla liczby wyselekcjonowanych cech

Importantly, the first feature in the ranking (the highest FeatureRank value) obtained the value of the measure equal to 1. This means that this feature always appeared in the first place in each of the 50 created models, i.e. the implemented algorithm each time, regardless of the selected training set, considered this feature as the most important, enabling the best division of the two classes in the set. Figure 3 also shows the cut-off point for the number of features selected for the model. Below this threshold, the features are characterized by insignificant differences in the value of the determined FeatureRank significance measure.

By first using the wrapper method with the use of a classifier based on logistic regression, and then the filter method based on the entered significance measure, the dimensionality of the data set was significantly reduced from 406 input features to 29 selected features.

Table 3

Parameter values for final model

Feature ID	**Intercept**	**1**	**2**	**3**	**4**
Parameter	-2.4687	0.7876	0.2519	-0.2097	0.7381
	5	**6**	**7**	**8**	**9**
	1.2674	0.7112	0.4770	-0.6276	-0.2917
	10	**11**	**12**	**13**	**14**
	0.8514	0.3900	-0.2118	0.3395	0.8966
	15	**16**	**17**	**18**	**19**
	-0.1732	-0.6903	-0.3346	-0.4697	-0.2658
	20	**21**	**22**	**23**	**24**
	-0.1970	-0.1940	-0.0599	-0.4010	-0.3079
	25	**26**	**27**	**28**	**29**
	0.2947	-0.2597	-0.2527	-0.4715	-0.1366

In the next step, the quality of selected features was assessed in a given research problem, i.e. the ability to separate elements from the control class and procedural class. For this purpose, parameter values were determined for individual features included in the final model, maintaining the order of features resulting from the obtained FeatureRank metric values. The values of the estimated parameters are shown in Table 3.

In order to determine the threshold value of the classification probability, the Youden index method was used. This method finds a trade-off between the sensitivity and specificity of the classification, and the computation is stepwise for each possible value of the classification probability. For this purpose, the ROC curve was determined and the value of the threshold classification probability was estimated as the point farthest from the diagonal of the plot. The new probability threshold value for the classification, determined with the use of the Youden index, was 0.7047.

In the last step of the study, the quality of the constructed classifier based on logistic regression methods was determined, based on the validation set. It is worth emphasizing that this set was not used at any stage of constructing the classifier and was completely randomly selected from the set of input data.

Table 4

Classification quality metrics based on the validation set

Quality metric	Result
PPV	0.9385
NPV	0.9147
Sensitivity	0.8977
Specificity	0.9491
Weighted accuracy	0.9253

The following measures of classification quality were used: Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity, Specificity, and the weighted quality of the classification, which takes into account the disproportions in the representation of individual classes Table 4 shows the results related to the classification quality of the validation set.

12.6. Discussion

The application of the proposed approach, taking into account the use of the forward selection technique as the leading method, requires large computational resources and is quite time-consuming, especially when considering large-scale data sets. Undoubtedly, a very big benefit resulting from the use of a time-consuming and computationally complicated methodology of forward selection on a full set of features is its potential to reveal often hidden relationships between successive features. By using filters, as the first FS method, we are able to significantly limit the input set of features to the most important ones, in other words allowing to capture the greatest differentiation between the studied groups. However, by using such a scheme, we can remove the hidden connections between the features in the first stage. It is particularly important when building classifiers to distinguish various biological complexities, in very heterogeneous data sets, related to gene expression, e.g. healthy tissue and cancer tissue, control cells and irradiated cells, distinguishing types of cancers, etc. In very complex biological problems, it is important to capture any differences, but also the connections and cooperation of individual features and their common, cumulative ability to evaluate a given phenomenon. Hence, the applied approach based on the use of wrapper techniques in the first phase of feature selection,

which is much more time-consuming and computationally complicated than filters, allowed for the capture of many hidden inter-features dependencies. The further methodology, i.e. the use of a simple metric to determine the significance of individual selected features, allowed for a further reduction in dimensionality and the selection of only the most essential features. The proposed method is not only able to capture what is seemingly invisible in the analyzed sets of features, but also allows you to select features that separate classes of elements present in the given classes with very satisfactory results, which was confirmed in this work. What is more, using the proposed integration methodology, we are able to build the final model that can be successfully used to classify elements from highly heterogeneous data sets.

12.7. Conclusions

The presented results concerning the classification quality measures for 29 selected out of 406 features unambiguously allow to state that the selected features allow, with satisfactory results, to separate the elements from the control and procedural groups. The constructed classifier is characterized by a very high specificity, i.e. the ability to correctly classify elements from the control group, and a slightly lower, but still high value of sensitivity, i.e. the ability to correctly classify elements from the procedural group. The very good results of the classification are clearly evidenced by the high value of the weighted quality at a level above 92.5%, while the 95% confidence interval for mean weighted accuracy value for 50 build models was equal (91.05÷91.61).

Referring to the applied FS methodology, it allowed for a significant reduction in the dimensionality of the data. The first stage of FS, i.e. the application of the method belonging to the wrapper group - a classifier using the forward selection procedure, allowed for the identification of 159 important features from the point of view of the classification problem. Already this stage allowed for a significant reduction in the number of features that were subjected to the second stage of FS, i.e. the use of the methodology from the group of filters. Describing each of the 159 features, using a measure based on the previously calculated values of the classification quality on the test sets and the order of attaching subsequent features to individual models, allowed for the selection of features that have a significant impact on the quality of the classification. At this stage, there was another dimensionality reduction from 159 to

29 features that were incorporated into the final model. What is also worth emphasizing, the final model for heterogeneous data set classification, was built from 29 features, while the 95% confidence interval for the average value of the number of features over the 50 built models was (28.51÷31.05).

The presented method, integrating the methodology of wrappers and filters, allows for a significant reduction in dimensionality (the number of features), while maintaining a very high quality of classification, in relation to a very heterogeneous set of data.

Acknowledgments

This work has been supported by European Union under the European Social Fund grant AIDA – POWR.03.02.00-00-I029, and by the Silesian University of Technology Grant BK-02/070/BK22/0033.

Bibliography

1. J. Grimmer, B.M. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts., Political analysis 21.3, pp. 267–297, (2013).
2. J. Ferreira, M.A.T. Figueiredo: Efficient feature selection filters for high-dimensional data., Pattern recognition letters 33.13, pp. 1794–1804, (2012).
3. R. Spencer, F. Thabtah, N. Abdelhamid, M. Thompson: Exploring feature selection and classification methods for predicting heart disease, Digital health 6, p. 2055207620914777, (2020).
4. L. Blum, P. Langley: Selection of relevant features and examples in machine learning, Artificial intelligence 97.1–2, pp. 245–271, (1997).
5. R. Kohavi, G.H. John: Wrappers for feature subset selection, Artificial intelligence 97.1–2, pp. 273–324, (1997).
6. M. Mera-Gaona, D.M. López, R. Vargas-Canas, U. Neumann: Framework for the Ensemble of Feature Selection Methods, Applied Sciences 11.17, p. 8122, (2021).
7. J. Tang, S. Alelyani, H. Liu: Feature selection for classification: A review, Data classification: Algorithms and applications, p. 37, (2014).
8. I. Guyon, A. Elisseeff, An introduction to variable and feature selection., Journal of machine learning research, pp. 1157–1182, (2003).

9. H. Liu, H. Motoda (ed.): Computational methods of feature selection, CRC Press, (2007).
10. L. Liu, J. Kang, J. Yu, Z. Wang: A comparative study on unsupervised feature selection methods for text clustering, in 2005 International Conference on Natural Language Processing and Knowledge Engineering. IEEE,(2005).
11. X. He, D. Cai, P. Niyogi: Laplacian score for feature selection, Advances in neural information processing systems 18, (2005).
12. Q. Gu, Z. Li, J. Han: Generalized fisher score for feature selection, arXiv preprint arXiv:1202.3725, (2012).
13. K. Kira, L.A. Rendell: A practical approach to feature selection, Machine learning proceedings 1992. Morgan Kaufmann, pp. 249–256, (1992).
14. M. Robnik-Šikonja, I. Kononenko: Theoretical and empirical analysis of ReliefF and RReliefF, Machine learning 53.1 , pp. 23–69, (2003).
15. L. Yu, H. Liu: Feature selection for high-dimensional data: A fast correlation-based filter solution, in Proceedings of the 20th international conference on machine learning (ICML-03), (2003).
16. H. Peng, C. Ding: Minimum redundancy and maximum relevance feature selection and recent advances in cancer classification, Feature Selection for Data Mining 52, (2005).
17. H. Liu, L. Yu: Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on knowledge and data engineering 17.4, pp. 491–502, (2005).
18. S. Ma, J. Huang: Penalized feature selection and classification in bioinformatics, Briefings in bioinformatics 9.5 , pp. 392–403, (2008).
19. Y. Saeys, I. Inza, L. Pedo: A review of feature selection techniques in bioinformatics, bioinformatics 23.19, pp. 2507–2517, (2007).
20. J.R. Quinlan, C4. 5: programs for machine learning, Elsevier, (2014).
21. M. Sandri, P. Zuccolotto: Variable selection using random forests, Springer, in Data analysis, classification and the forward search., Berlin, Heidelberg, (2006), 263–270.
22. G. Cawley, N. Talbot, M. Girolami: Sparse multinomial logistic regression via bayesian l1 regularisation, in Advances in neural information processing systems 19, (2006).
23. C. Nguyen, Y. Wang, H.N. Nguyen: Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic, Journal of Biomedical Science and Engineering, (2013).

24. S. Ma, J. Huang: Penalized feature selection and classification in bioinformatics, Briefings in bioinformatics 9.5 , pp. 392–403, (2008).
25. A. Jović, K. Brkić, N. Bogunović: A review of feature selection methods with applications 2015, in 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), Ieee, (2015).
26. I. Guyon, J. Weston, S. Barnhill, V. Vapnik: Gene selection for cancer classification using support vector machines, Machine learning 46.1, pp. 389–422, (2002).
27. R. Tibshirani: Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58.1, pp. 267–288, (1996).
28. V. Fonti, E. Belitser: Feature selection using lasso, VU Amsterdam research paper in business analytics 30 , pp. 1–25, (2017).
29. V.N. Vapnik, A.Y. Chervonenkis: Necessary and sufficient conditions for the uniform convergence of means to their expectations, Theory of Probability & Its Applications 26.3, pp. 532–553, (1982).
30. R.-C. Chen, C. Dewi, S.-W. Huang, R.E. Caraka: Selecting critical features for data classification based on machine learning methods, Journal of Big Data 7.1, pp. 1–26, (2020).
31. R. Quinlan, in C4.5: Programs for Machine Learning. Morgan Kaufmann, (1993).
32. J.R. Quinlan: Induction of decision trees. Machine learning, 1(1), (1986), p. 81–106.

FEATURE SELECTION METHODS FOR CLASSIFICATION PURPOSES

Abstract

Feature selection methods are nowadays more and more developed. Modern and very accurate techniques that allow for the generation of very extensive data sets are also becoming more popular. Text analysis sections and the new, promising single-cell sequencing technique are specific areas with the privilege of high-dimensional data. A very large amount of expenditure on the continuous improvement of feature selection methods is widely appreciated by scientists and analysts, especially now. Methods combining two popular techniques of feature selection – wrappers and filters – are becoming more and more popular. The method proposed by us, combining the

effectiveness of wrappers techniques and the speed of filters techniques, allows you to choose the features important from the classification point of view with great efficiency. These features are capable of carrying significant information about the differences between elements coming from different classes. What is more, our method also allows us to capture many invisible, without complicated analyzes, relationships between the analyzed features. The effectiveness of the proposed methodology is supported by a very high quality of heterogeneous data set classification at the level above 92.5%, as well as very satisfactory sensitivity and specificity metrics.

Keywords: feature selection, logistic regression, machine learning, heterogeneous data sets classification