Feature Selection Methods Based on Decision Rule and Tree Models

Wiesław Paja

Abstract Feature selection methods, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. In this work, a novel concepts of relevant feature selection based on information gathered from decision rule and decision tree models were introduced. A new measures *DRQualityImp* and *DTLevelImp* were additionally defined. The first one is based on feature presence frequency and rule quality, while the second is based on feature presence on different levels inside decision tree. The efficiency and effectiveness of that method is demonstrated through the exemplary use of five real-world datasets. Promising initial results of classification efficiency could be gained together with substantial reduction of problem dimensionality.

Keywords Feature selection · Feature ranking · Decision rules · Dimensionality reduction · Relevance and irrelevance

1 Introduction

In the era of the acquisition of vast amounts of data, different domain information databases, efficient analysis and retrieval of regularity has become an extremely important task. The issue of classification and object recognition is applied in many fields of human activity. Data mining is fraught with many aspects which hinder it like a very large number of observations, too many attributes, the insignificance of the part of variables for the classification process, mutual interdependence of conditional variables, the simultaneous presence of variables with different types, the

W. Paja (✉)
Faculty of Mathematics and Natural Sciences, Department of Computer Science,
University of Rzeszów, 1 Prof. S. Pigonia Street, 35-310 Rzeszów, Poland
e-mail: wpaja@ur.edu.pl

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies 2016*,
Smart Innovation, Systems and Technologies 57,
DOI 10.1007/978-3-319-39627-9_6

presence of undefined values of variables, the presence of erroneous values of the variables, uneven distribution of categories for the target variable. Thus, the development of efficient methods for significant feature selection is valid.

This kind of methods are frequently used as a preprocessing steps to machine learning experiments. It could be defined as a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion. Feature selection has been a fruitful field of research and development since 1970s and proven to be effective in removing irrelevant features, increasing efficiency in learning tasks, improving learning performance like predictive accuracy, and enhancing comprehensibility of learned results [1].

The feature selection methods are typically presented in three classes based on how they combine the selection algorithm and the model building: filter, wrapper and embedded FS methods. Filter methods select features regardless of the model. They are based only on general features like the correlation with the variable to predict. These methods select only the most interesting variables. Then selected subset will be part of a classification model. Such methods are effective in computation time and robust to overfitting [2]. But, some redundant, but relevant, features are not recognized. In turn, wrapper methods evaluate subsets of features which allows to detect the possible interactions between variables [1, 3, 4]. However, the increasing overfitting risk when the number of observations is insufficient could be possible. Additionally, the significant computation time when the number of variables is large highly increase. The third type called embedded methods devotes to reduce the classification of learning. These methods try to combine the advantages of both previous methods. Thus, the learning algorithm takes advantage of its own variable selection algorithm. So, it needs to know initially what a good selection is, which limits their exploitation [5].

Kohavi and John [1] observed that there are several definitions of relevance that may be contradictory and misleading. They proposed that two degrees of relevance (*strong* and *weak*) are required to encompass all notions that are usually associated with this term. In their approach the relevance is defined in the absolute terms, with the help of ideal Bayes classifier. In this context a feature X is *strongly relevant* when removal of X alone from the data always results in deterioration of the prediction accuracy of the ideal Bayes classifier. In turn, feature X is *weakly relevant* if it is not strongly relevant and there exists a subset of features S, such that the performance of ideal Bayes classifier on S is worse than the performance on $S \cup \{X\}$. A feature is *irrelevant* if it is neither strongly or weakly relevant.

Nilsson and co-workers [6] introduced the formal definition of two different feature selection problems *Minimal Optimal Feature Selection (MOSF)* and *All Relevant Feature Selection (ARFS)*. *MOSF* means identification of minimal set of features to obtain optimum quality classification. In turn, *ARFS* devotes to find all the variables that may, under certain conditions, improve the classification. There are two important differences between these problems. The first one is detection of attributes with low importance (*ARFS*) [7], which may be completely obscured by other, more important attributes from the point of view of the classifier (*MOFS*). The second difference is to find the boundary between the variables poorly, but realistically related to the decision and those for whom such a relation is created as a result of random fluctuations. The formal definition of the problem of all relevant feature selection (*ARFS*) as a distinct problem from the classical minimal optimal feature selection (*MOFS*), was proposed as recently as 2007 [6].

Here, two different wrapper methods of feature importance calculation are presented. The first one apply frequency of each feature occurrence inside rules, and the second method comply decision tree structure during calculation. Similar methods for selecting and evaluating most important rule features based on the rule accuracy, frequency of the elementary condition in the discovered rule set, and its influence for the quality of the whole set of generated rules was extensively analyzed in [8]. Another approach [9] devotes to method for evaluating the importance of gene ontology terms which compose multi-attribute rules. The obtained ranking is used to generate a new set of rules that provide additional information about the biological function of genes.

Additionally, in this research, to distinguish between relevant and irrelevant features the *contrast variable* concept [7] were applied. It is a variable that does not carry information on the decision variable by design that is added to the system in order to discern relevant and irrelevant variables. Here, it is obtained from the real variables by random permutation of values between objects. The use of contrast variables was for the first time proposed by Stoppiglia and co-workers [10] and then by Tuv and co-workers [11].

2 Methods and Algorithms

During experiments the following general procedure was applied:

1. *Step: Selection of dataset and features for investigation (10-folds)*
    - *Addition of contrast features to original data*
    - *Application of a set of ranking measures to calculate importance for each feature*
    - *Calculation of DRQualityImp (or DTLevelImp) importance parameter for each feature*
    - *Definition (selection) of the most important feature subset*
2. *Step: Application of different machine learning algorithms for classification of unseen objects (the same 10-folds like in Step 1)*
    - *Using all original features*
    - *Using only selected, important features*
3. *Step: Comparison of gathered results using evaluation measures*

In the first step, dataset and feature for investigation were defined. Different ranking measures were applied to estimate importance of each feature. In order to check specificity of the feature selection, the dataset was extended by adding contrast variables. It means that each original variable was duplicated and its values were randomly permuted between all objects. Hence a set of non-informative by design shadow variables was added to original variables. The variables that were selected as important significantly than random, were examined further, using different test. To define level of feature importance six well-known ranking measures were applied: *ReliefF, Information Gain, Gain Ratio, Gini Index, SVM weight* and *RandomForest*. Additionally, new measures based on decision rules and decision trees, called *DRQualityImp* and *DTLevelImp*, were introduced. The first one is based on the frequency of presence of the different feature in rule model that is generated from original dataset and also takes into consideration the quality of rules in which this feature occurs. Thus, the *DRQualityImp* of the *ith* attribute could be presented in the Eq. 1.

$$DRQualityImp_{A_i} = \sum_{j=1}^{n} Q_{R_j}\{A_i\} \tag{1}$$

where n is a number of rules inside the model, Q_{R_j} defines classification quality of the rule R_j and A_i describe the presence of the *ith* attribute, usually *1* (feature occurred) or *0* (feature didnt occur). In turn, quality of rule is defined in the Eq. 2.

$$Q_{R_j} = \frac{E_{corr}}{E_{corr} + E_{incorr}} \tag{2}$$

where E_{corr} depicts the number of correctly matched learning objects by the *jth* rule and E_{incorr} depicts the number of incorrectly matched learning objects by this rule.

Moreover, second new measure, based on decision trees, called *DTLevelImp* were defined. It is based on the presence of different feature in the decision tree nodes generated from original dataset and also takes into consideration the product of weight W_j assigned to a given level j of the tree and the number of cases *Inst(node)* classified in a given *node* at this level in which feature A_i occurs. Thus, the *DTLevelImp* of the *ith* attribute could be presented in the Eq. 3.

$$DTLevelImp_{A_i} = \sum_{j=1}^{l} \sum_{node=1}^{x} W_j * Inst(node) * \{A_i\} \tag{3}$$

where l is the number of levels inside the model, x is the number of nodes inside given level and A_i describe the presence of the *ith* attribute, usually *1* (feature occurred) or *0* (feature didnt occur). In turn, weight W of level j is defined in the Eq. 4.

$$W_j = \begin{cases} 1 & j = 1, j \in N \\ \frac{W_{j-1}}{2} & 1 < j \leq l \end{cases} \tag{4}$$

During the second step the test probing the importance of variables was performed by analyzing the influence of variables used for model building on the prediction quality.

Table 1 Summary characteristic of benchmark datasets

Dataset	# Instances	# Features	# Classes
Breast cancer	286	9	2
Heart disease	303	13	2
Lung cancer	32	56	3
Primary tumor	339	17	21
Skin cancer	548	13	4

Six different machine learning algorithms were applied to build different predictors for the original set of features and for selected features: *Classification Tree (CT), Random Forest (RF), CN2 decision rules algorithm (CN2), Naive Bayes (NB), k Nearest Neighbors (kNN)* and *Support Vector Machine (SVM)*. During this step the 10-fold cross validation paradigm were also applied using the same folds as it was in the first step. Nine known evaluation measures were applied in each predictor: *Classification Accuracy (CA), Sensitivity, Specificity, Area Under ROC curve (AUC), Information Score (IS), F1 score (F1), Precision, Brier measure* and *Matthew Coefficient Correlation (MCC)* [12]. Finally, two of them were summarized in Table 4.

3 Investigated Datasets

Initial investigations were focused on applying developed algorithms on several real-world datasets. Five datasets have been used during experiments. Four of them are gathered from UCI ML repository, and the fifth set have been developed earlier by the author [13]. The summary of datasets is presented in Table 1. These datasets have diverse number of objects, features and their types and also classes.

4 Results and Conclusions

To illustrate proposed methodology only results for Breast cancer datasets will be presented in details. The first step of the experiment revealed three features, that were recommended as important by all, or nearly all, ranking measures. In Table 2, we can observe that *deg-malig, node-caps,* and *irradiat* features create stable and core set of features which have the highest rank values using most of eight measures of importance, particularly using *DRQualityImp* measure, introduced in this investigation. In the same table, comparison with importance of contrast values (*"contrast"* index) is also presented. The most important contrast feature is *irradiat (contrast)* for which *DRQualityImp* measure is equal to *4.59*. In this way, it is also treated as a threshold that separates the core, relevant set of attributes from other less infor-

Table 2 Ranking of features using eight different measures

Feature	ReliefF	Inf. gain	Gain ratio	Gini	SVM weight	RF	DRQuality Imp	DTLevel Imp
deg-malig	−0.02	**0.08**	**0.05**	**0.02**	0.07	**2.03**	**8.06**	**235.46**
node-caps	**0.03**	**0.06**	**0.08**	**0.02**	0.06	**1.98**	**7.94**	24.94
irradiat	**0.01**	0.03	**0.03**	0.01	0.02	**0.69**	**5.64**	3.78
irradiat (contrast)	−0.05	0	0	0	0.05	**0.12**	**4.59**	5.21
inv-nodes	**0.03**	**0.07**	**0.05**	**0.02**	0.06	0.07	4.52	17.62
breast	−0.08	0	0	0	0.02	**0.33**	3.66	3.44
menopause (contrast)	**−0.01**	0.01	0.01	0	0.07	−0.01	3.44	9.82
menopause	−0.06	0	0	0	0.03	0	3.21	8.85
node-caps (contrast)	−0.02	0	0	0	0.03	0.07	2.78	21.01
inv-nodes (contrast)	−0.05	0.02	**0.01**	0	**0.17**	−0.02	2.39	12.42
breast-quad (contrast)	−0.12	0.01	0	0	0.06	−0.02	2.10	13.43
deg-malig (contrast)	−0.07	0	0	0	0.01	0	1.89	3.71
age (contrast)	−0.11	0.02	0.01	0	0.14	0.1	1.85	21.08
breast (contrast)	−0.06	0	0	0	0.03	0.1	1.71	3.45
breast-quad	−0.11	0.01	0.01	0	0.13	0.1	1.48	**45.50**
tumor-size	−0.13	**0.06**	**0.02**	**0.01**	0.1	0.01	1.32	**58.38**
tumor-size (contrast)	−0.16	**0.03**	0.01	**0.01**	0.11	−0.01	0.88	**29.6**
age	−0.1	0.01	0.01	0	0.05	0.06	0	7.36

mative attributes. Most of the measures (except *SVM weight*) used in this approach show that selected set of features has higher values of these parameters than gathered threshold value (underlined values). These values are denoted in bold style in Table 2. Hereby, we can observe that different measures give different threshold.

It should be stressed that using the *DTLevelImp* parameter the selected set of features is different: *deg-malig, breast-quad* and *tumor-size*. This selected set of features achieve results of classification similar to original one (see Table 4). Thus, some redundant information could be recognized.

The second step of experiment devoted to evaluation of prediction quality of utilized machine learning algorithms described in Sect. 2. During this step six different algorithms were applied using 10-fold cross validation method. Average results for the *Breast cancer* dataset are collected in Table 3. Three types of results is presented: achieved using original dataset, achieved using the cuted set of features by application of *DRQualityImp* and *DTLevelImp* indicator.

This procedure was applied to two specified sets:

- the original dataset containing all descriptive features,
- the dataset containing only selected features according to their importance calculated in the first step.

Finally, all average results for *Breast cancer* dataset are collected in Table 3. Based on these results, it could stressed that set of selected features which contains only *3* from *9* attributes has similar (even better) prediction quality (*CA* and *AUC*)

Table 3 Average results of classification quality for the *Breast cancer* dataset

Model	CA	Sens	Spec	AUC	IS	F1	Prec	Brier	MCC
On original data									
CT	0.68	0.57	0.57	0.57	0.00	0.59	0.58	0.49	0.14
CN2	0.74	0.61	0.61	0.71	0.06	0.61	0.72	0.37	0.30
SVM	0.75	0.61	0.61	0.68	0.04	0.65	0.76	0.37	
RF	0.76	0.60	0.60	0.69	0.03	0.67	0.78	0.37	
kNN	0.73	0.63	0.63	0.65	0.15	0.63	0.68	0.46	0.31
NB	0.74	0.67	0.67	0.69	0.12	0.67	0.69	0.43	0.36
On data selected using *DRQualityImp*									
CT	0.74	0.61	0.61	0.69	0.08	0.65	0.69	0.37	0.29
CN2	0.75	0.64	0.64	0.70	0.08	0.67	0.74	0.36	
SVM	0.76	0.62	0.62	0.66	0.06	0.65	0.78	0.38	
RF	0.76	0.62	0.62	0.71	0.06	0.65	0.78	0.37	
kNN	0.70	0.62	0.62	0.61	0.00	0.62	0.67	0.43	0.28
NB	0.75	0.66	0.66	0.72	0.11	0.66	0.73	0.37	0.38
On data selected using *DTLevelImp*									
CT	0.73	0.63	0.63	0.63	0.02	0.63	0.67	0.41	0.30
CN2	0.68	0.53	0.53	0.66	−0.01	0.63	0.61	0.38	
SVM	0.70	0.51	0.51	0.68	−0.04	0.80	0.69	0.39	
RF	0.71	0.57	0.57	0.69	0.00	0.70	0.61	0.38	
kNN	0.72	0.63	0.63	0.66	0.07	0.62	0.68	0.44	0.29
NB	0.72	0.63	0.63	0.66	0.03	0.66	0.64	0.40	0.26

Table 4 Summary results of feature selection and classification

Dataset	Measure	Original set	DRQualityImpset	DTLevelImp set
Breast cancer	#Features	**9**	**3** (33.3 %)	**3** (33.3 %)
	CA	0.73 ± 0.03	0.74 ± 0.02	0.71 ± 0.02
	AUC	0.67 ± 0.05	0.68 ± 0.04	0.66 ± 0.04
Heart disease	#Features	**13**	**8** (61.5 %)	**6** (46.2 %)
	CA	0.80 ± 0.03	0.79 ± 0.03	0.80 ± 0.02
	AUC	0.87 ± 0.05	0.86 ± 0.04	0.88 ± 0.04
Lung cancer	#Features	**56**	**3** (5.4 %)	**4** (7.1 %)
	CA	0.52 ± 0.09	0.52 ± 0.09	0.53 ± 0.07
	AUC	0.70 ± 0.05	0.70 ± 0.05	0.73 ± 0.05
Skin cancer	#Features	**13**	**8** (61.5 %)	**9** (69.2 %)
	CA	0.82 ± 0.02	0.79 ± 0.01	0.79 ± 0.02
	AUC	0.96 ± 0.02	0.95 ± 0.01	0.95 ± 0.01
Primary tumor	#Features	**17**	**13** (76.5 %)	**12** (70.6 %)
	CA	0.42 ± 0.04	0.42 ± 0.03	0.42 ± 0.04
	AUC	0.83 ± 0.04	0.83 ± 0.04	0.83 ± 0.04

as it was observed with all original attributes. Furthermore, all other measures in Table 3 also increased a little. With the exception of *Brier* score, which decreased, but the lower the *Brier* score is for a set of predictions, the better the predictions are calibrated [14].

Similar results were obtained for other investigated datasets (see Table 4). All number of features in selected sets are significantly less than in original one. It is average about *45 %* of original features selected. Using these selected sets promising initial results of classification efficiency could be gained together with substantial reduction of problem dimensionality.

Acknowledgments This work was supported by the Center for Innovation and Transfer of Natural Sciences and Engineering Knowledge at the University of Rzesżw.

References

1. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
2. Bermingham, M.L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., Agakov, F., Navarro, P., Haley, C.S.: Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. **5**, (2015)
3. Phuong, T.M., Lin, Z., Altman, R.B.: Choosing SNPs using feature selection. In: Proceedings of 2005 IEEE Computational Systems Bioinformatics Conference, CSB 2005, pp. 301–309 (2005)
4. Paja, W., Wrzesien, M., Niemiec, R., Rudnicki, W.R.: Application of all-relevant feature selection for the failure analysis of parameter-induced simulation crashes in climate models. Geosci. Model Dev. **9**, 1065–1072 (2016)
5. Zhu, Z., Ong, Y.S., Dash, M.: Wrapper-filter feature selection algorithm using a memetic framework. IEEE Trans. Syst. Man, Cybern. Part B Cybern. **37**, 70–76 (2007)
6. Nilsson, R., Peña, J.M., Björkegren, J., Tegnér, J.: Detecting multivariate differentially expressed genes. BMC Bioinf. **8**, 150 (2007)
7. Rudnicki, W.R., Wrzesień, M., Paja, W.: All Relevant feature selection methods and applications. In: Stańczyk, U., Lakhmi, C.J. (eds.) Feature Selection for Data and Pattern Recognition, pp. 11–28. Springer-Verlag, Berlin Heidelberg, Berlin (2015)
8. Greco, S., Słowinski, R., Stefanowski, J.: Evaluating importance of conditions in the set of discovered rules. In: RSFDGrC'07: Proceedings of the 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Toronto, Ontario, Canada, pp. 314–321 (2007)
9. Sikora, M., Gruca, A.: Quality improvement of rules based gene groups descriptions using information about GO terms importance occurring in premises of determined rules. Int. J. Appl. Math. Comput. Sci. **20**(3), 555–570 (2010)
10. Stoppiglia, H., Dreyfus, G., Dubois, R., Oussar, Y.: Ranking a random feature for variable and feature selection. J. Mach. Learn. Res. **3**, 1399–1414 (2003)
11. Tuv, E., Borisov, A., Torkkola, K.: Feature selection using ensemble based ranking against artificial contrasts. In: International Symposium on Neural Networks, pp. 2181–2186 (2006)
12. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**, 861–874 (2006)
13. Hippe, Z.S., Bajcar, S., Blajdo, P., Grzymala-Busse, J.P., Grzymala-Busse, J.W., Knap, M., Paja, W., Wrzesien, M.: Diagnosing skin melanoma: current versus future directions. TASK Q. **7**, 289–293 (2003)
14. Hernández-Orallo, J., Flach, P., Ferri, C.: A unified view of performance metrics: translating threshold choice into expected classification loss. J. Mach. Learn. Res. **13**, 2813–2869 (2012)