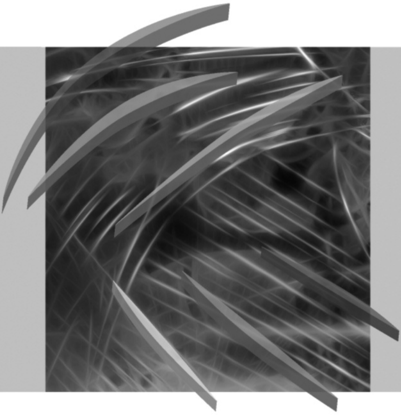
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

206

Advanced Information Technologies for Management – AITM 2011

Intelligent Technologies and Applications



edited by
**Jerzy Korczak, Helena Dudycz,
Mirosław Dyczkowski**

Publishing House of Wrocław University of Economics
Wrocław 2011

Reviewers: Frederic Andres, Witold Chmielarz, Jacek Cypryjański, Beata Czarnacka-Chrobot,
Bernard F. Kubiak, Halina Kwaśnicka, Antoni Ligęza, Anna Ławrynowicz,
Mikołaj Morzy, Stanisław Stanek, Ewa Ziemba

Copy-editing: Agnieszka Flasińska

Layout: Barbara Łopusiewicz

Proof-reading: Marcin Orszulak

Typesetting: Adam Dębski

Cover design: Beata Dębska

This publication is available at www.ibuk.pl

Abstracts of published papers are available in the international database
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
and in The Central and Eastern European Online Library www.ceeol.com

Information on submitting and reviewing papers is available on the Publishing House's website
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
ISBN 978-83-7695-182-9

The original version: printed
Printing: Printing House TOTEM

Contents

Streszczenia

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 206
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS

Advanced Information Technologies for Management – AITM 2011
Intelligent Technologies and Applications

ISSN 1899-3192

Paweł Ziemba, Mateusz Piwowarski*
West Pomeranian University of Technology, Szczecin, Poland

FEATURE SELECTION METHODS IN DATA MINING TECHNIQUES

Abstract: Data mining techniques are largely based on machine learning algorithms. They are to serve to extract data models which, due to their large information content, are not recognized by people. Data redundancy poses a problem both for data mining algorithms as well as people, which is why various methods are used in order to reduce the amount of analyzed data, including data mining methods such as feature selection. The article outlines basic issues linked with feature selection and contains an analysis of five feature selection algorithms belonging to the filter category. Results obtained by each method were validated with the help of CART decision tree algorithms. The CART analysis revealed that the results of each of the five algorithms are acceptable.

Keywords: data mining, dimension reduction, feature selection, feature filters.

1. Introduction

Data mining techniques are used to extract patterns from data sets which, due to the extent of the analyzed data, are not recognized by people. Data mining methods are mainly based on machine learning algorithms. Machine learning tasks concentrate on predicting an object's value or class affiliation, based on its features. The multidimensionality of an object which is to be classified into a specific category poses a problem for classification techniques, as well as for all data mining methods. Dimensionality is a serious obstacle, impacting the effectiveness of data mining algorithms and the machine learning methods they utilize, as the amount of data that must be analyzed with the help of data mining algorithms increases considerably when numerous dimensions are involved. This problem is referred to as the *curse of dimensionality* [Chizi, Maimon 2010]. A reduction in the number of dimensions undergoing classification allows for a reduction of calculation demands and data collection demands, as well as the increased reliability of predicate results and data quality [Guyon 2008]. Dimension reduction can be conducted with the help of two methods.

[*] e-mails: {pziemba, mpiwowarski}@wi.zut.edu.pl.

1. With the help of a feature extraction process which involves the extraction of a set of new characteristics from the original features. This process usually involves remapping the original features in a way that creates new variables. Factor analysis is an example of this type of dimension reduction.

2. With the help of a feature selection process concentrated on pinpointing significant features within the data set and rejecting redundant attributes. Various evaluation algorithms are used to assess features according to specific criteria which describe their significance for the classification task [Hand, Mannila, Smyth 2005].

The feature selection process can be described as searching through a set of characteristics describing an object undergoing classification, according to specific assessment criteria. The process entails two procedures: filtering and wrapping. Filters are based on independent feature assessments, using general data characteristics. Feature sets are filtered in order to establish the most promising attribute subset before commencing machine learning algorithm training [Witten, Frank 2005]. Wrapper functions evaluate specific feature subsets with the help of machine learning algorithms. The learning algorithm is, in this case, included in the feature selection procedure [Hall, Holmes 2003]. Each of these procedures contains four elements:

1) generating a feature subset,
2) subset assessment,
3) stop criterion,
4) result validation [Liu, Yu, Motoda 2003].

Basic subset generating procedures include: creating an individual ranking, backwards search and forward search. The individual ranking procedure does not take into account the dependencies between attributes – it analyzes each object feature individually – due to which its results can be less reliable than the results obtained using the remaining strategies described here. Backwards and forward searches are greedy strategies which give suboptimal results [Michalak, Kwaśnicka 2006]. Subset assessment is conducted with the help of filtering or wrapping methods. Feature subset testing is, however, always conducted with the use of a machine learning algorithm [Hsu, Hsieh, Lu 2011].

Wrapper functions differ from one another only in terms of utilized machine learning algorithms, so the results obtained with the help of wrappers depend only on the quality of the machine learning algorithm and whether or not it suits the given classification task. Due to the above, this article concentrates only on the analysis of the feature subset evaluation algorithm, used during filtration procedures where features are assessed with the help of means other than degree of correct classification criteria.

2. Selection methods based on filters

The following feature selection procedures utilize methods which involve filters:

– ReliefF,

- – LVF (Las Vegas Filter),
- – FCBF (Fast Correlation Based Filter),
- – CFS (Correlation-based Feature Selection),
- – SA (Significance Attribute).

The primary idea behind the ReliefF method is to evaluate attributes according to how well differentiate between similar objects, i.e. objects which have similar feature values. The nearest neighbors method is used here; a proximity function [Kononenko 1994]. The ReliefF procedure utilizes a heuristics according to which a good attribute should differentiate objects situated close to each other but belonging to various classes, and additionally should maintain the same value for objects situated close to each other but belonging to the same class. The method of establishing an evaluation for an X_i feature can be expressed with the help of formula (1) [Kononenko, Hong 1997]:

$$q(X_i)=\sum_{r}\sum_{C(s)\neq C(r)}\left(\frac{P(C(s))}{1-P(C(r))}*\frac{d_{rs,i}}{k}\right)-\sum_{r}\sum_{C(s)=C(r)}\frac{d_{rs,i}}{k}. \qquad (1)$$

The proximity function is expressed with the help of formula (2) [Kira, Rendell 1992]:

$$d_{rs,i}=\begin{cases} 0 \quad \text{when} \quad r(i)=s(i) \\ 1 \quad \text{when} \quad r(i)\neq s(i) \\ \begin{array}{c} r(i)-s(i) \\ nu(i) \end{array} \end{cases} \quad \text{for} \quad \text{nominal} \quad \text{attributes} \qquad (2)$$

The LVF method utilizes the probabilistic approach in order to establish the direction of the correct solution. Solution searches are conducted randomly, which guarantees an acceptable solution even if incorrect decisions are made during the search for the best subset. The method uses the inconsistency criterion to determine the level of acceptance of data with reduced dimensionality. The inconsistency coefficient can be expressed with the help of formula (3) [Liu, Setiono 1996]:

$$IncR=\frac{\sum_{i}D_i-M_i}{N}, \qquad (3)$$

where D_i expresses the number of occurrences of the *i*-th feature value combination, M_i expresses the number of objects in the dominating class for the *i*-th combination of attributes and N expresses the number of objects.

The FCBF method is based on the correlation coefficient, or, more precisely, symmetrical uncertainty. Symmetrical uncertainty is defined as the relationship of the informational content of a pair of attributes to the entropy sum of these attributes, and is expressed by formula (4) [Kannan, Ramaraj 2010]:

$$SU(X,Y) = 2 * \left[\frac{IG(X \mid Y)}{H(X) + H(Y)} \right]. \tag{4}$$

Additionally, as auxiliary means, the FCBF method utilizes sets of redundant features, separately for each feature. The Sp_i^+ set contains redundant features for the F_i feature, with a higher symmetrical uncertainty coefficient than F_i, in relation to class C, whilst the Sp_i^- set contains redundant features for the F_i feature, with a lower symmetrical uncertainty coefficient, in relation to class C. The FCBF procedure initially involves calculation of the symmetrical uncertainty for each feature, and further considerations involve only attributes with symmetrical uncertainty values higher than the assumed threshold. These are then added to the *S'* set in descending order, based on symmetrical uncertainty values. Set *S'* is then analyzed for the presence of redundant features [Yu, Liu 2003].

The CFS method, like the FCBF method, is based on an analysis of the correlations between features. The global correlation measure used by the CFS procedure is Pearson's linear correlation, whilst symmetrical uncertainty is utilized as a local measure. CFS uses a heuristics stating that a good feature subset contains attributes strongly correlated with a specific class of objects but uncorrelated with other classes and attributes [Hall, Smith 1998]. The CFS method utilizes formula (5) [Hall, Smith 1999]:

$$Merit_s = \frac{k * r_{cf}}{\sqrt{k + k * (k-1) * r_f}}, \tag{5}$$

where *Merit* is the value of the heuristic for subset *S* containing *k* features, r_{cf} is the average value of the correlation coefficient between features for subset *S* and object classes, whilst r_{ff} expresses the average mutual correlation between features. The CFS heuristic filters out features which describe the affiliation of an object to a class only to a small degree as well as redundant features strongly correlated with other features [Hall, Smith 1999]. The CFS method initially maps a mutual correlation matrix between attributes and the correlation between attributes and classes of objects with the help of symmetrical uncertainty calculations. Once this step is completed, the *first best* forward search algorithm is employed [Hall 2000].

The attribute significance method utilizes the bidirectional link coefficient to assess links between attributes and class affiliation. This method is based on a heuristic stating that if an attribute is significant, then there is a big probability that objects complementing value sets for its attributes will belong to complementary class sets. Additionally assuming that decision classes for two objects sets differ, it can be expected that the attribute significance value for objects belonging to two different sets will also differ. The significance of each attribute is expressed as the average value of general links: of a given attribute with classes (AE) and classes with a given attribute (CE) [Ahmad, Dey 2005].

3. Research procedure

The research procedure involved generating feature rankings with the help of each of the above-mentioned methods. The generated rankings were then validated with the use of a classifier. The classifier was used to assess feature sets from which the least significant attributes for each of the obtained rankings were eliminated with the use of iteration. Research was to indicate differences in rankings obtained with the help of each of the methods and to determine the smallest attribute sets which allow for the correct classification of objects and decrease the level of data redundancy. The objective was thus to determine features which were in actuality significant for the data sets. CART decision trees were used as classifiers, utilizing G-square measures and a-priori decision class affiliation probability evaluations, depending on the amount of objects in each class in the training set [Rokach, Maimon 2010a; Webb 2003]. A 10-fold cross validation was employed to stabilize classification results [Rokach, Maimon 2010b].

Three datasets from the UCI Machine Learning Repository [UCI] website were analyzed: the Car Evaluation Data Set, Image Segmentation Data Set and Wine Quality Data Set [Cortez et al. 2009]. The Car Evaluation set contained 1728 objects, of which each was described with the help of 6 attributes with discrete values, and could belong to one of 4 classes determining car purchase admissibility; each class contained a different number of objects. The Image Segmentation set contained 2100 objects described with the help of 19 attributes. Each object could belong to one of 7 classes expressing the content of a graphic image described by an object. Each class contained the same number of objects, and the attributes used had constant values. The Wine Quality set expressed affiliation of white wines to one of 10 quality classes. The set contained 11 constant attributes and 4898 objects, variously distributed among the specific classes. Such a selection of data sets allowed to examine work of individual algorithms in the situation, when sets are characterized different: with the cardinality of decision-making classes, the cardinality of conditional attributes and the degree of the membership of objects in individual decision-making classes. In the selection of data sets (in the perspective of the authors' further works) a fact that two of sets described quality classes of examined objects was important.

Choice of methods of features selection, which was applied at the work, was not also random. Every of studied features selection methods (except for CFS and FCBF methods) is characterized by a different approach towards selection and uses other heuristics. CFS and FCBF methods also differ from each other in some respects, and the examination allowed to determine to what extent the differences between these methods affect the results. LVF and CFS methods used forward searches as attribute subset generating strategies, whilst the remaining three methods utilized the individual ranking strategy. In the case of the ReliefF method, 10 closest neighbors were used to assess attributes, whilst sampling was conducted for all objects.

4. Research results

Significance rankings obtained with the help of each of the above-mentioned methods for specific datasets are presented in Tables 1, 2 and 3.

Table 1. Feature significance for the Car Evaluation set

ReliefF	Feature	6	4	1	2	5	3
	Importance	0.3573	0.2908	0.2195	0.1944	0.0371	–0.0535
LVF	Feature	1	6	4	2	5	3
	Inconsistency	0.7	0.703	0.819	0.892	0.962	1
FCBF	Feature	6	4	1	2	5	3
	Importance	0.1879	0.1574	0.0602	0.046	0.0215	0.0028
CFS	Feature	6	4	1	2	5	3
	Importance	0.1879	0.1727	0.1352	0.1129	0.0946	0.0793
SA	Feature	6	4	1	2	5	3
	Importance	0.4334	0.3846	0.2455	0.2049	0.119	0.0567

Table 2. Feature significance for the Image Segmentation set

ReliefF	Feature	20	13	18	11	12	14	3	17	16	15	19	2	9	7	5	6	10	8	4
	Importance	0.2205	0.2201	0.2161	0.202	0.1961	0.1945	0.1904	0.1728	0.1618	0.1457	0.1451	0.0651	0.029	0.0231	0.0145	0.0077	0.0037	0.002	0
LVF	Feature	12	3	20	7	2	10	16	9	4	5	6	8	11	13	14	15	17	18	19
	Inconsistency	0.661	0.893	0.974	0.985	0.99	0.994	0.996	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
FCBF	Feature	12	20	11	14	13	18	17	19	3	16	15	9	7	10	8	2	6	5	4
	Importance	0.5629	0.5568	0.5212	0.5044	0.5026	0.501	0.4433	0.4322	0.4305	0.4186	0.3735	0.1828	0.175	0.1533	0.1375	0.0519	0.0153	0.0123	0
CFS	Feature	12	20	3	14	17	19	15	13	9	16	11	2	7	18	6	5	10	8	4
	Importance	0.563	0.677	0.705	0.709	0.71	0.711	0.709	0.705	0.7	0.695	0.692	0.689	0.684	0.68	0.675	0.67	0.664	0.656	0.545
SA	Feature	20	12	11	18	14	19	13	17	3	16	15	7	9	8	10	6	2	5	4
	Importance	0.9637	0.959	0.9407	0.9343	0.9317	0.9069	0.9019	0.8843	0.8828	0.8363	0.7828	0.5655	0.5644	0.5076	0.5043	0.2849	0.2519	0.1755	0

The same order in the ranking for the Car Evaluation set was obtained by each method, with the exception of the LVF procedure. The LVF method gave different results from the remaining procedures for the first three positions. As LVF utilizes incoherence coefficients to assess criteria significance it is the only researched method which presented feature orders in an ascending order (most significant features had the lowest incoherence values).

Bigger differences can be observed in the case of Image Segmentation and Wine Quality sets. The LVF procedure results varied most for these sets as well. In the case of the CFS method it can clearly be seen that the calculated significances do not reflect accurately the order in the feature significance ranking. Rankings for this procedure are corrected with the use of the *best first* strategy, and only then do they resemble rankings obtained with the help of the remaining methods (with the exception of the LVF method). The large similarity between the CFS and FCBF methods causes a feature with the highest position in rankings formed with the use of these two methods to always have an identical or nearly identical significance value.

Table 3. Feature significance for the Wine Quality set

ReliefF	Feature	11	2	9	10	1	7	3	4	6	5	8
	Importance	0.0166	0.0111	0.0103	0.0093	0.0084	0.0083	0.0082	0.0066	0.0064	0.0046	0.0041
LVF	Feature	11	2	4	3	6	7	10	5	8	9	1
	Inconsistency	0.493	0.539	0.569	0.617	0.679	0.743	0.805	0.854	0.888	0.91	0.915
FCBF	Feature	11	8	5	7	3	6	2	4	9	1	10
	Importance	0.09	0.0652	0.0488	0.0351	0.0347	0.0338	0.0324	0.0318	0.0117	0.0115	0.009
CFS	Feature	11	8	5	3	2	6	7	4	1	9	10
	Importance	0.09	0.0975	0.1026	0.106	0.1089	0.112	0.1118	0.1108	0.109	0.1073	0.1053
SA	Feature	11	4	6	5	8	2	3	7	1	9	10
	Importance	0.3545	0.3089	0.3043	0.2625	0.259	0.2484	0.2299	0.2234	0.1527	0.1334	0.0961

Ranking verification results are presented in Tables 4, 5 and 6. Results for the classification of objects in the Car Evaluation set, presented in Table 4, with the use of smaller and smaller feature subsets, are similar for all methods. Differences appear only for feature subsets containing two attributes; in this case the attribute subset generated by the LVF method has a correct classification coefficient 7% lower than in the case of the remaining methods. Only objects belonging to two of four classes were correctly classified. For subsets containing amounts of 3, correct classification is for three out of four classes.

Table 4. Correlation of the % of correct classifications after the removal of following features for the Car Evaluation set

ReliefF FCBF CFS SA	Feature	6	4	1	2	5	3
	mean	70.02	77.78	81.94	89.24	96.18	97.45
	min	0	0	0	30.43	86.96	88.41
LVF	Feature	1	6	4	2	5	3
	mean	70.02	70.25	81.94	89.24	96.18	97.45
	min	0	0	0	30.43	86.96	88.41

In case of the Image Segmentation data set (classification results are presented in Table 5), the best classification results are obtained with the help of feature subsets generated by the LVF algorithm. Although in the case of subsets containing larger amounts of attributes, results are more accurate with the use of any of the remaining methods, especially ReliefF and CFS, the LVF procedure handles smaller feature subsets the best. Good results are also obtained with the help of the CFS algorithm.

Results obtained by the remaining methods, although of lesser quality, were nonetheless acceptable.

Table 5. Correlation of the % of correct classifications after the removal of following features for the Image Segment set

ReliefF	Feature	20	13	18	11	12	14	3	17	16	15	19	2	9	7	5	6	10	8	4
	mean	68.38	87.86	87.86	89.43	90.62	91.10	96.95	97.00	97.00	97.00	96.90	97.71	98.05	98.05	98.05	98.05	98.10	98.10	98.10
	min	38.33	69.67	69.67	57.00	72.00	73.00	88.67	88.67	88.67	88.67	88.33	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00
LVF	Feature	12	3	20	7	2	10	16	9	4	5	6	8	11	13	14	15	17	18	19
	mean	69.10	92.76	96.90	96.90	97.67	97.67	97.76	97.76	97.76	97.76	97.76	97.76	97.81	97.81	97.81	97.81	97.81	97.81	98.10
	min	29.33	78.33	88.33	88.33	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.33	94.33	94.33	94.33	94.33	94.33	94.00
FCBF	Feature	12	20	11	14	13	18	17	19	3	16	15	9	7	10	8	2	6	5	4
	mean	69.10	89.00	89.57	89.62	91.10	91.10	92.29	92.81	96.90	96.90	96.90	97.43	97.43	97.48	97.48	98.10	98.10	98.10	98.10
	min	29.33	69.33	68.00	70.67	73.00	73.00	77.33	80.00	88.33	88.33	88.33	93.00	93.00	93.00	93.00	94.00	94.00	94.00	94.00
CFS	Feature	12	20	3	14	17	19	15	13	9	16	11	2	7	18	6	5	10	8	4
	mean	69.10	89.00	96.90	96.90	96.95	96.81	96.81	96.86	97.38	97.38	97.43	98.05	98.05	98.05	98.05	98.05	98.10	98.10	98.10
	min	29.33	69.33	88.33	88.33	88.33	88.00	88.00	88.33	93.00	93.00	93.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00
SA	Feature	20	12	11	18	14	19	13	17	3	16	15	7	9	8	10	6	2	5	4
	mean	68.38	89.00	89.57	89.43	91.10	92.71	92.71	92.81	96.90	96.90	96.90	96.90	97.43	97.43	97.48	97.48	98.10	98.10	98.10
	min	38.33	69.33	68.00	57.00	73.00	78.00	78.00	80.00	88.33	88.33	88.33	88.33	93.00	93.00	93.00	93.00	94.00	94.00	94.00

Results obtained for the Wine Quality set are presented in Table 6. Based on the obtained data it is difficult to assess which method offers best results, as results depend on the size of the feature subset. If the feature subset is to be reduced by one attribute, the best results are obtained by eliminating feature number 5, indicated by the ReliefF function. The elimination of this feature lowers the correct classification coefficient only by 0.06%, but simultaneously improves results for the third worst-recognized decision class (classification results for the two worst-recognized classes are not included in Table 6 as these results oscillated from 0 to 10%). As subsequent features are eliminated, these methods offer various results, yet within acceptable levels.

Table 6. Correlation of the % of correct classifications after the removal of following features for the Wine Quality set

ReliefF	Feature	11	2	9	10	1	7	3	4	6	5	8
	mean	50.98	59.74	62.19	64.07	64.01	64.82	65.31	65.43	66.99	67.99	68.05
	min	0	11.66	15.95	18.40	20.25	18.40	21.47	17.71	33.71	27.61	26.99
LVF	Feature	11	2	4	3	6	7	10	5	8	9	1
	mean	50.98	59.74	63.05	62.84	62.47	63.39	63.80	65.21	65.72	66.78	68.05
	min	0	11.66	19.63	13.14	21.47	30.06	27.43	20.86	25.77	24.54	26.99
FCBF	Feature	11	8	5	7	3	6	2	4	9	1	10
	mean	50.98	58.49	60.45	60.25	62.17	63.70	63.09	64.52	65.01	65.78	68.05
	min	0	1.23	6.75	10.86	18.29	18.86	16.57	24.54	23.31	29.45	26.99
CFS	Feature	11	8	5	3	2	6	7	4	1	9	10
	mean	50.98	58.49	60.45	62.37	64.90	63.21	63.09	64.52	65.27	65.78	68.05
	min	0	1.23	6.75	16.56	19.63	11.43	16.57	24.54	26.99	29.45	26.99
SA	Feature	11	4	6	5	8	2	3	7	1	9	10
	mean	50.98	58.76	59.82	62.09	63.37	64.23	64.07	64.52	65.27	65.78	68.05
	min	0	6.13	16.56	22.09	20.86	22.70	22.86	24.54	26.99	29.45	26.99

5. Conclusions

The article outlines the problem of multivariate data, for which data mining techniques are used to find patterns not seen by people. Both people and data mining techniques must deal with the problem posed by the massive amount of data which must be analyzed. People have problems with determining patterns even if the dataset is not very big. Data mining techniques can deal with the problem, yet data redundancy can lead to the decreased quality of the identified patterns and an increase in the amount of time required for data mining algorithms to analyze data. The significant feature selection method is one of the methods used to reduce the amount of data analyzed via data mining. The analysis described in the article pertained to the assessment of five feature selection methods, with the help of distance, correlation and probability criteria, used to analyze three datasets. The study determined feature assessment means used by each of the methods and included a validation of the attribute assessment results with the help of a machine learning algorithm, the CART decision tree.

Research results show a relatively large similarity between feature assessment conducted with the help of the FCBF and CFS procedures. This is due to the procedural similarities shared by the methods: both methods utilize correlation measures, or, strictly speaking, symmetrical uncertainty. The LVF method, which utilizes the incoherence criterion, gave results which differed most from results obtained with the help of the remaining methods. The CFS method may give rise to some objections, as here the significance coefficient did not exactly reflect the order of features in rankings formed with the help of this method. In result these means cannot be used e.g. as significances for a given criterion in decision facilitating tasks. As far as the quality of features obtained with the use of each of methods is concerned, or the accuracy of choice of significant features, validation results illustrate that each method generates acceptable features. Accuracy of classification results generated by each of the methods depends on the number of attribute subsets. It cannot be thus stated that one method is significantly superior to another in feature selection; the number of expected feature sets must always be taken under consideration. The dataset for which selection is to take place also plays a significant role, as results show that various methods bring various results, depending on the classification tasks. It is possible to generalize these conclusions for similar sets of data, like the ones which were exploited at the work.

References

Ahmad A., Dey L. (2005), A feature selection technique for classificatory analysis, *Pattern Recognition Letters*, Vol. 26, pp. 43–56.

Chizi B., Maimon O. (2010), Dimension reduction and feature selection, [in:] O. Maimon, L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook*, Springer, New York, pp. 83–100.

Cortez P., Cerdeira A., Almeida F., Matos T., Reis J. (2009), Modeling wine preferences by data mining from physicochemical properties, *Decision Support Systems*, Vol. 47, No. 4, pp. 547–553.

Guyon I. (2008), Practical feature selection: From correlation to causality, [in:] F. Fogelman-Soulié, D. Perrotta, J. Piskorski, R. Steinberger (Eds.), *Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and Their Applications to Security*, IOS Press, Amsterdam, pp. 27–43.

Hall M.A. (2000), Correlation-based feature selection for discrete and numeric class machine learning, [in:] *ICML'00 Proceedings of the 17th International Conference on Machine Learning*, pp. 359–366.

Hall M.A., Holmes G. (2003), Benchmarking attribute selection techniques for discrete class data mining, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 3, pp. 1437–1447.

Hall M.A., Smith L.A. (1998), Practical feature subset selection for machine learning, [in:] *Proceedings of Australasian Computer Science Conference*, pp. 181–191.

Hall M.A., Smith L.A. (1999), Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper, [in:] *Proceedings of the 12th International Florida Artificial Intelligence Research Society Conference*, pp. 235–239.

Hand D., Mannila H., Smyth D. (2005), *Eksploracja danych*, WNT, Warszawa, pp. 414–416.

Hsu H., Hsieh C., Lu M. (2011), Hybrid feature selection by combining filters and wrappers, *Expert Systems with Applications*, Vol. 38, pp. 8144–8150.

Kannan S.S., Ramaraj N. (2010), A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm, *Knowledge-Based Systems*, Vol. 23, pp. 580–585.

Kira K., Rendell L.A. (1992), A practical approach to feature selection, [in:] *ML92 Proceedings of the 9th International Workshop on Machine Learning*, pp. 249–256.

Kononenko I. (1994), Estimating attributes: Analysis and extensions of RELIEF, *Lecture Notes in Computer Science*, Vol. 784, pp. 171–182.

Kononenko I., Hong S.J. (1997), Attribute selection for modelling, *Future Generation Computer Systems*, Vol. 13, No. 2–3, 1997, pp. 181–195.

Liu H., Setiono R. (1996), A probabilistic approach to feature selection – A filter solution, *The 13th*

International Conference on Machine Learning ICML'96, pp. 319–327.

Liu H., Yu L., Motoda H. (2003), Feature extraction, selection, and construction, [in:] N. Ye (Ed.), *The Handbook of Data Mining*, Lawrence Erlbaum Associates, Mahwah, pp. 409–424.

Michalak K., Kwaśnicka H. (2006), Correlation-based feature selection strategy in classification problems, *International Journal of Applied Mathematics and Computer Science*, Vol. 16, No. 4, pp. 503–511.

Rokach L., Maimon O. (2010a), Classification trees, [in;] O. Maimon, L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook*, Springer, New York, pp. 149–174.

Rokach, L., Maimon, O. (2010b), Supervised learning, [in:] O. Maimon, L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook*, Springer, New York, pp. 133–148.

UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/index.html

Webb G.I. (2003), Association rules, [in:] N. Ye (Ed.), *The Handbook of Data Mining*, Lawrence Erlbaum Associates, Mahwah, pp. 25–40.

Witten I.H., Frank E. (2005), *Data Mining. Practical Machine Learning Tools and Techniques*, Elsevier, San Francisco, pp. 288–295.

Yu L., Liu H. (2003), Feature selection for high-dimensional data: A fast correlation-based filter solution, [in:] *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, , pp. 856–863.

METODY SELEKCJI CECH W TECHNIKACH *DATA MINING*

Streszczenie: Techniki *data mining* w większości oparte są na algorytmach uczenia maszynowego. Służą one wykrywaniu w danych wzorców, które z powodu bardzo dużej ilości informacji są niewidoczne dla człowieka. Jednak dla algorytmów *data mining*, podobnie jak dla człowieka, problemem jest nadmiarowość danych. W związku z tym stosowane są metody mające na celu redukcję ilości danych analizowanych przez metody *data mining*, takie jak np. selekcja cech. W artykule omówiono podstawowe zagadnienia związane z zagadnieniem selekcji cech. Przybliżono i zbadano działanie pięciu algorytmów selekcji cech, należących do kategorii filtrów. Walidacja wyników selekcji wykonanej za pomocą każdej z metod została wykonana z użyciem algorytmu drzew decyzyjnych CART. Uzyskane rezultaty wskazują na akceptowalność wyników otrzymanych z użyciem każdej z badanych metod.

Słowa kluczowe: *data mining*, redukcja ilości danych, selekcja cech, filtry cech.