A review of feature selection methods with applications

A. Jović*, K. Brkić* and N. Bogunović*
* Faculty of Electrical Engineering and Computing, University of Zagreb / Department of Electronics, Microelectronics, Computer and Intelligent Systems, Unska 3, 10 000 Zagreb, Croatia
{alan.jovic, karla.brkic, nikola.bogunovic}@fer.hr

Abstract - Feature selection (FS) methods can be used in data pre-processing to achieve efficient data reduction. This is useful for finding accurate data models. Since exhaustive search for optimal feature subset is infeasible in most cases, many search strategies have been proposed in literature. The usual applications of FS are in classification, clustering, and regression tasks. This review considers most of the commonly used FS techniques. Particular emphasis is on the application aspects. In addition to standard filter, wrapper, and embedded methods, we also provide insight into FS for recent hybrid approaches and other advanced topics.

I. INTRODUCTION

The abundance of data in contemporary datasets demands development of clever algorithms for discovering important information. Data models are constructed depending on the data mining tasks, but usually in the areas of classification, regression and clustering. Often, pre-processing of the datasets takes place for two main reasons: 1) reduction of the size of the dataset in order to achieve more efficient analysis, and 2) adaptation of the dataset to best suit the selected analysis method. The former reason is more important nowadays because of the plethora of developed analysis methods that are at the researcher's disposal, while the size of an average dataset keeps growing both in respect to the number of features and samples.

Dataset size reduction can be performed in one of the two ways: feature set reduction or sample set reduction. In this paper, the focus is on feature set reduction. The problem is important, because a high number of features in a dataset, comparable to or higher than the number of samples, leads to model overfitting, which in turn leads to poor results on the validation datasets. Additionally, constructing models from datasets with many features is more computationally demanding [1]. All of this leads researchers to propose many methods for feature set reduction. The reduction is performed through the processes of feature extraction (transformation) and feature selection. Feature extraction methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Multidimensional Scaling work by transforming the original features into a new feature set constructed from the original one based on their combinations, with the aim of discovering more meaningful information in the new set [2]. The new feature set can then be easily reduced by taking into consideration characteristics such as dataset variance coverage. Feature selection, on the other hand, is a process of taking a small subset of features from the original feature set without transformation (thus preserving the interpretation) and validating it with respect to the analysis goal. The selection process can be achieved in a number of ways depending on the goal, the resources at hand, and the desired level of optimization.

In this paper, we focus on feature selection and provide an overview of the existing methods that are available for handling several different classes of problems. Additionally, we consider the most important application domains and review comparative studies on feature selection therein, in order to investigate which methods perform best for specific tasks. This research is motivated by the fact that there is an abundance of work in this field and insufficient systematization, particularly with respect to various application domains and novel research topics.

Feature set reduction is based on the terms of feature relevance and redundancy with respect to goal. More specifically, a feature is usually categorized as: 1) strongly relevant, 2) weakly relevant, but not redundant, 3) irrelevant, and 4) redundant [3,4]. A strongly relevant feature is always necessary for an optimal feature subset; it cannot be removed without affecting the original conditional target distribution [3]. Weakly relevant feature may not always be necessary for an optimal subset, this may depend on certain conditions. Irrelevant features are not necessary to include at all. Redundant features are those that are weakly relevant but can be completely replaced with a set of other features such that the target distribution is not disturbed (the set of other features is called Markov blanket of a feature). Redundancy is thus always inspected in multivariate case (when examining feature subset), whereas relevance is established for individual features. The aim of feature selection is to maximize relevance and minimize redundancy. It usually includes finding a feature subset consisting of only relevant features.

In order to ensure that the optimal feature subset with respect to goal concept has been found, feature selection method has to evaluate a total of $2^m - 1$ subsets, where m is the total number of features in the dataset (an empty

feature subset is excluded). This is computationally infeasible even for a moderately large *m*. Therefore, putting completeness of the search aside, many heuristic methods have been proposed to find a sufficiently good (but not necessarily optimal) subset. The whole process of finding the feature subset typically consists of four basic steps: 1) subset generation, 2) subset evaluation, 3) a stopping criterion, and 4) validation of the results [5]. Feature subset generation is dependent on the state space search strategy. After a strategy selects a candidate subset, it will be evaluated using an evaluation criterion in step 2. After repeating steps 1 and 2 for a number of times depending on the process stopping criterion, the best candidate feature subset is selected. This subset is then validated on an independent dataset or using domain knowledge, while considering the type of task at hand.

II. Classification of Feature Selection Methods

Feature selection methods can be classified in a number of ways. The most common one is the classification into filters, wrappers, embedded, and hybrid methods [6]. The abovementioned classification assumes feature independency or near-independency. Additional methods have been devised for datasets with structured features where dependencies exist and for streaming features [2].

A. *Filter methods*

Filter methods select features based on a performance measure regardless of the employed data modeling algorithm. Only after the best features are found, the modeling algorithms can use them. Filter methods can rank individual features or evaluate entire feature subsets. We can roughly classify the developed measures for feature filtering into: information, distance, consistency, similarity, and statistical measures. While there are many filter methods described in literature, a list of common methods is given in Table I, along with the appropriate references that provide details. Not all the filter features can be used for all classes of data mining tasks. Therefore, the filters are also classified depending on the task: classification, regression or clustering. Due to lack of space, we do not consider semi-supervised learning feature selection methods in this work. An interested reader is referred to [16] for more information.

Univariate feature filters evaluate (and usually rank) a single feature, while multivariate filters evaluate an entire feature subset. Feature subset generation for multivariate filters depends on the search strategy. While there are many search strategies, there are four usual starting points for feature subset generation: 1) forward selection, 2) backward elimination, 3) bidirectional selection, and 4) heuristic feature subset selection. Forward selection typically starts with an empty feature set and then considers adding one or more features to the set. Backward elimination typically starts with the whole feature set and considers removing one or more features from the set. Bidirectional search starts from both sides - from an empty set and from the whole set, simultaneously considering larger and smaller feature subsets. Heuristic selection generates a starting subset based on a heuristic (e.g. a genetic algorithm), and then explores it further.

The most common search strategies that can be used with multivariate filters can be categorized into exponential algorithms, sequential algorithms and randomized algorithms. Exponential algorithms evaluate a number of subsets that grows exponentially with the feature space size. Sequential algorithms add or remove features sequentially (one or few), which may lead to local minima. Random algorithms incorporate randomness into their search procedure, which avoids local minima [17]. Common search strategies are shown in Table II.

TABLE I. Common Filter Methods for Feature Selection

Name	Filter class	Applicable to task	Study
Information gain	univariate, information	classification	[6]
Gain ratio	univariate, information	classification	[7]
Symmetrical uncertainty	univariate, information	classification	[8]
Correlation	univariate, statistical	regression	[8]
Chi-square	univariate, statistical	classification	[7]
Inconsistency criterion	multivariate, consistency	classification	[9]
Minimum redundancy, maximum relevance (mRmR)	multivariate, information	classification, regression	[2]
Correlation-based feature selection (CFS)	multivariate, statistical	classification, regression	[7]
Fast correlation-based filter (FCBF)	multivariate, information	classification	[8]
Fisher score	univariate, statistical	classification	[10]
Relief and ReliefF	univariate, distance	classification, regression	[11]
Spectral feature selection (SPEC) and Laplacian Score (LS)	univariate, similarity	classification, clustering	[4]
Feature selection for sparse clustering	multivariate, similarity	clustering	[12]
Localized Feature Selection Based on Scatter Separability (LFSBSS)	multivariate, statistical	clustering	[13]
Multi-Cluster Feature Selection (MCFS)	multivariate, similarity	clustering	[4]
Feature weighting K-means	multivariate, statistical	clustering	[14]
ReliefC	univariate, distance	clustering	[15]

TABLE II. Search Strategies for Feature Selection

Algorithm group	Algorithm name
Exponential	Exhaustive search Branch-and-bound
Sequential	Greedy forward selection or backward elimination Best-first Linear forward selection Floating forward or backward selection Beam search (and beam stack search) Race search
Randomized	Random generation Simulated annealing Evolutionary computation algorithms (e.g. genetic, ant colony optimization) Scatter search

B. Wrapper methods

Wrappers consider feature subsets by the quality of the performance on a modelling algorithm, which is taken as a black box evaluator. Thus, for classification tasks, a wrapper will evaluate subsets based on the classifier performance (e.g. Naïve Bayes or SVM) [18,19], while for clustering, a wrapper will evaluate subsets based on the performance of a clustering algorithm (e.g. *K*-means) [20]. The evaluation is repeated for each subset, and the subset generation is dependent on the search strategy, in the same way as with filters. Wrappers are much slower than filters in finding sufficiently good subsets because they depend on the resource demands of the modelling algorithm. The feature subsets are also biased towards the modelling algorithm on which they were evaluated (even when using cross-validation). Therefore, for a reliable generalization error estimate, it is necessary that both an independent validation sample and another modelling algorithm are used after the final subset is found. On the other hand, it has been empirically proven that wrappers obtain subsets with better perfomance than filters because the subsets are evaluated using a real modelling algorithm. Practically any combination of search strategy and modelling algorithm can be used as a wrapper, but wrappers are only feasible for greedy search strategies and fast modelling algorithms such as Naïve Bayes [21], linear SVM [22], and Extreme Learning Machines [23].

C. Embedded and hybrid methods

Embedded methods perform feature selection during the modelling algorithm's execution. These methods are thus embedded in the algorithm either as its normal or extended functionality. Common embedded methods include various types of decision tree algorithms: CART, C4.5, random forest [24], but also other algorithms (e.g. multinomial logistic regression and its variants [25]). Some embedded methods perform feature weighting based on regularization models with objective functions that minimize fitting errors and in the mean time force the feature coefficients to be small or to be exact zero. These methods based on Lasso [26] or Elastic Net [27] usually work with linear classifiers (SVM or others) and induce penalties to features that do not contribute to the model.

Hybrid methods were proposed to combine the best properties of filters and wrappers. First, a filter method is used in order to reduce the feature space dimension space, possibly obtaining several candidate subsets [28]. Then, a wrapper is employed to find the best candidate subset. Hybrid methods usually achieve high accuracy that is characteristic to wrappers and high efficiency characteristic to filters. While practically any combination of filter and wrapper can be used for constructing the hybrid methodology, several interesting methodologies were recently proposed, such as: fuzzy random forest based feature selection [29], hybrid genetic algorithms [30], hybrid ant colony optimization [31], or mixed gravitational search algorithm [32].

D. Structured and streaming features

In some datasets, features may exhibit certain internal structures such as spatial or temporal smoothness, disjoint/overlapping groups, tree- or graph-like structures. In these datasets, features are not independent. Therefore, it is a good idea to employ specific algorithms to deal with the dependencies in order to increase performance of the selected feature subsets. Most of the algorithms dealing with feature structures are recent, and are based on some adaptation of the Lasso regularization method to accomodate different structures. Good overviews of these methods can be found in [2,33].

Streaming (or dynamic) features are features which size is unknown in advance; they are rather dynamically generated, they arrive as streamed data into the dataset and the modelling algorithms has to reach a decision whether to keep them as useful for model construction or not. Also, some features may become irrelevant over time and should be discarded. This scenario is common in social networks such as Twitter, where new words are generated that are not all relevant for a given subject [2]. The most important feature selection methods in this category are: the Grafting algorithm [34], the Alpha-Investing algorithm [35] the OSFS algorithm [36], and dynamic feature selection fuzzy-rough set approach [37].

III. Feature Selection Application Domains

The choice of feature selection methods differs among various application areas. In the following subsections, we review comparative studies on feature selection pertaining to several well known application domains. Table III summarizes the findings from the reviewed studies.

A. Text mining

In text mining, the standard way of representing a document is by using the bag-of-words model. The idea is to model each document with the counts of words occurring in that document. Feature vectors are typically formed so that each feature (i.e. each element of the feature vector) represents the count of a specific word, an alternative being to just indicate the presence/absence of a word without specifying the count. The set of words whose occurrences are counted is called a vocabulary. Given a dataset that needs to be represented, one can use all the words from all the documents in the dataset to build the vocabulary and then prune the vocabulary using feature selection.

It is common to apply a degree of preprocessing prior to feature selection, typically including the removal of rare words with only a few occurrences, the removal of overly common words (e.g. "a", "the", "and" and similar) and grouping the differently inflected forms of a word together (lemmatization, stemming) [38].

Forman [38] performed a detailed experimental study of filter feature selection methods for text classification. Twelve feature selection metrics were evaluated on 229 text classification problem instances. Feature vectors were formed not as word counts, but as Boolean representations of whether a certain word occurred or not. A linear SVM classifier with untuned parameters was used to evaluate performance. The results were analyzed with respect to precision, recall, F-measure and accuracy. Information gain was shown to perform best

TABLE III. SUMMARIZED FINDINGS OF RELEVANT FEATURE SELECTION METHODS IN VARIOUS APPLICATION AREAS

Application area	Subfield	Datasets	Feature selection methods	Evaluation metrics	Best performing	Study
Text mining	Text classification	229 text classification problem instances gathered from Reuters, TREC, OHSUMED, etc.	Accuracy, accuracy balanced, bi-normal separation, chi-square, document frequency, F1-measure, information gain, odds ratio, odds ration numerator, power, probability ratio, random	Accuracy, F-measure, precision, and recall	Information gain (precision), bi-normal separation (accuracy, F-measure, recall)	[38]
	Text clustering	Reuters-21578, 20 Newsgroups, Web Directory	Information gain, chi-square, document frequency, term strength, entropy-based ranking, term contribution, iterative feature selection	Entropy, precision	Iterative feature selection	[39]
Image processing / computer vision	Image classification	Aerial Images, The Digits Data, Cats and Dogs	Relief (R), K-means (K), sequential floating forward selection (F), sequential floating backward selection (B), various combinations R + K + F/B	Average MSE of 100 neural networks	R+K+B / R+K+F / R+K, depending on the size of feature subset	[40]
	Breast density classification from mammographic images	Mini-MIAS, KBD-FER	Best-first with forward, backward and bi-directional search, genetic search and random search (k-NN and Naïve Bayesian classifiers)	Accuracy	Best first forward, best first backward	[41]
Bioinformatics	Biomarker discovery	Three benchmark datasets deriving from DNA microarray experiments	Chi-square, information gain, symmetrical uncertainty, gain ratio, OneR, ReliefF, SVM-embedded	Stability, AUC	Chi-square, symmetrical uncertainty, information gain, ReliefF	[42]
	Microarray gene expression data classification	Two gene expression datasets (Freije, Phillips)	Information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, t-statistics, one-dimensional SVM	Accuracy	Consensus of all methods	[43]
Industrial applications	Fault diagnosis	Wind turbine test rig dataset	Distance, entropy, SVM wrapper, neural network wrapper, global geometric similarity scheme	Accuracy	Global geometric similarity scheme with wrapper	[22]

with respect to precision, while the author-introduced method bi-normal separation performed best for recall, F-measure and accuracy.

Liu et al. [39] investigated the use of feature selection in the problem of text clustering, showing that feature selection can improve its performance and efficiency. Five filter feature selection methods were tested on three document datasets. Unsupervised feature selection methods were shown to improve clustering performance, achieving about 2% entropy reduction and 1% precision improvement on average, while removing 90% of the features. The authors also proposed an iterative feature selection method inspired by expectation maximization that combines supervised feature selection methods with clustering in a bootstrap setting. The proposed method reduces the entropy by 13.5% and increases precision by 14.6%, hence coming closest to the established baseline, obtained by using a supervised approach.

B. Image processing and computer vision

Representing images is not a straightforward task, as the number of possible image features is practically unlimited [40]. The choice of features typically depends on the target application. Examples of features include histograms of oriented gradients, edge orientation histograms, Haar wavelets, raw pixels, gradient values, edges, color channels, etc. [44].

Bins and Draper [40] studied the use of filter feature selection methods in the general problem of image classification. Three different image datasets were used. They proposed a three-step method for feature selection that combines Relief, K-means clustering and sequential floating forward/backward feature selection (SFFS/SFBS). The idea is to: 1) use the Relief algorithm to remove irrelevant features, 2) use K-means clustering to cluster similar features and remove redundancy, and 3) run SFFS or SFBS to obtain the final set of features. The authors found that using the proposed hybrid combination of algorithms yields better performance than when using Relief or SFFS/SFBS alone. In cases when there are no irrelevant or redundant features in the dataset, the proposed algorithm does not degrade performance. When the goal is to select a specific number of features, it is suggested to use the R+K+B variant of the algorithm if the number of relevant and non-redundant features is less than 110, and otherwise R+K+F. If the number of selected features is allowed to vary, authors suggest using

R+K. The authors also note that Relief is good at removing irrelevant features, but not adequate for selecting the best among relevant features.

Muštra et al. [41] investigated the use of wrapper feature selection methods for breast density classification in mammographic images. Five wrapper feature selection methods were evaluated in conjunction with three different classifiers on two datasets of mammographic images. The best-performing methods were best-first search with forward selection and best-first search with backward selection. Overall, the results over different classifiers and datasets were improved between 3% and 12% when using feature selection.

C. Industrial applications

Feature selection is important in fault diagnosis in industrial applications, where numerous redundant sensors monitor the performance of a machine. Liu et al. [22] have shown that the accuracy of detecting a fault (i.e. solving a binary classification problem of machine state as faulty vs. normal) can be improved by using feature selection. They proposed to use a global geometric model and a similarity metric for feature selection in fault diagnostics. The idea is to find feature subsets that are geometrically similar to the original feature set. The authors experimented with three different similarity measures: angular similarity, mutual information and structure similarity index. The proposed approach was compared with distance-based and entropy-based feature selection, and with SVM and neural network wrappers. The best performance was obtained by combining the proposed geometric similarity approach with a wrapper, so that top 10% of feature subsets were preselected by geometric similarity, following by an exhaustive search-based wrapper approach to find the best subset.

D. Bioinformatics

An interesting application of feature selection is in biomarker discovery from genomics data. In genomics data, individual features correspond to genes, so by selecting the most relevant features, one gains important knowledge about the genes that are the most discriminative for a particular problem. Dessì et al. [42] proposed a framework for comparing different biomarker selection methods, taking into account predictive performance and stability of the selected gene sets. They compared eight selection methods on three benchmark datasets derived from DNA microarray experiments. Additionally, they analyzed how similar the outputs of different selection methods are, and found that the outputs of univariate methods seem to be more similar to each other than to the multivariate methods. In particular, the SVM-embedded selection seems to select features quite distinct from the ones selected by other methods. When jointly optimizing stability and predictive performance, best results were obtained using chi- square, systematic uncertainty, information gain and ReliefF.

Abusamra [43] analyzed the performance of eight different filter-based feature selection methods and three classification methods on two datasets of microarray gene expression data. The best individually performing feature selection methods varied depending on the dataset and the classifier used. Notably, using Gini index for feature selection improved performance of an SVM classifier on both datasets. Some feature selection methods were shown to degrade classification performance. However, Abusamra demonstrated that classification accuracy can be consistently improved on both datasets using a *consensus* of all feature selection methods to find top 20 features, by counting the number of feature selection methods that selected each feature. Seven features were selected by all the methods, and additional 13 features were randomly selected from a pool of features selected by seven out of eight methods.

IV. Conclusion

The current research advancement in this field is identified in the area of hybrid feature selection methods, particularly concerning the methodologies based on evolutionary computation heuristic algorithms such as swarm intelligence based and various genetic algorithms. Additionally, application areas such as bioinformatics, image processing, industrial applications and text mining deal with high-dimensional feature spaces where a clever hybrid methodology design is of utmost importance if any success is to be obtained. Therein, features may exhibit complex internal structures or may even be unknown in advance.

While there is no silver bullet method, filters based on information theory and wrappers based on greedy stepwise approaches seem to offer best results. Future research should focus on optimizing the efficiency and accuracy of feature subset search strategy by combining earlier best filter and wrapper approaches. Most research tends to focus on small number of datasets on which their methodology works. Larger comparative studies should be pursued in order to have more reliable results.

Acknowledgements

This work has been supported in part by the Croatian Science Foundation, within the project “De-identification Methods for Soft and Non-Biometric Identifiers” (DeMSI, UIP-11-2013-1544). This support is gratefully acknowledged.

References

[1] F. Korn, B. Pagel, and C. Faloutsos, "On the „dimensionality curse‟ and the „self-similarity blessing‟," *IEEE Trans. Knowl. Data Eng.* vol. 13, no. 1, pp. 96–111, 2001.

[2] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," in: C. Aggarwal (ed.), Data Classification: Algorithms and Applications. CRC Press, 2014.

[3] L Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," J. Mach. Learn. Res., vol. 5, pp. 1205–1224, 2004.

[4] S. Alelyani, J. Tang, and H. Liu, "Feature Selection for Clustering: A Review," in: C. Aggarwal and C. Reddy (eds.), Data Clustering: Algorithms and Applications, CRC Press, 2013.

[5] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, 2005.

[6] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method", *Expert Systems with Applications*, vol. 41, issue 14, pp. 6371–6385, 2014.

[7] I. H. Witten and E. Frank, Data mining: Practical machine learning tools and techniques, San Francisco CA, USA: Morgan Kaufmann, 2011.

[8] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in: *Proc. 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, USA, AAAI Press, pp. 856–863, 2003.

[9] H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection-A Filter Solution," in: *Proc. 13th International Conference on Machine Learning (ICML-1996)*, Bary, Italy, Morgan Kaufmann, pp. 319–327, 1996.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification, Wiley-interscience, 2012.

[11] M. R. Sikonja and I. Kononenko, "Theoretical and empirical analysis of Relief and ReliefF," *Mach. Learn.*, vol. 53, pp. 23–69, 2003.

[12] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 713–726, 2010.

[13] Y. Li, M. Dong, and J. Hua. "Localized feature selection for clustering," *Pattern Recognition Letters*, vol. 29, no. 1, pp. 10–18, 2008.

[14] D. S. Modha and W.S. Spangler, "Feature weighting in k-means clustering," *Mach. Learn.*, vol. 52, no. 3, pp. 217–237, 2003.

[15] M. Dash and Y.-S. Ong, "RELIEF-C: Efficient Feature Selection for Clustering over Noisy Data," in: *Proc. 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Roca Raton, Florida, USA, pp. 869–872, 2011.

[16] Z. Xu, I. King, and M. R.-T. Lyu, "Discriminative Semi-Supervised Feature Selection Via Manifold Regularization," *IEEE Trans. Neural Networks*, vol. 21, no. 7, pp. 1033–1047, 2010.

[17] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, London: Kluwer Academic Publishers, 1998.

[18] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in: *Proc. 15th International Conference on Machine Learning (ICML-1998)*, Madison, Wisconsin, USA, Morgan Kaufmann, pp. 82–90, 1998.

[19] S. Maldonado, R. Weber, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *Information Sciences*, vol. 286, pp. 228–246, 2014.

[20] Y. S. Kim, W. N. Street, and F. Menczer, "Evolutionary model selection in unsupervised learning," *Intelligent Data Analysis*, vol. 6, no. 6, pp. 531–556, 2002.

[21] J. C. Cortizo and I. Giraldez, "Multi Criteria Wrapper Improvements to Naive Bayes Learning," LNCS, vol. 4224, pp. 419–427, 2006.

[22] C. Liu, D. Jiang, and W. Yang, "Global geometric similarity scheme for feature selection in fault diagnosis," *Expert Systems with Applications*, vol. 41, issue 8, pp. 3585–3595, 2014.

[23] F. Benoît, M. van Heeswijk, Y. Miche, M. Verleysen, and A. Lendasse, "Feature selection for nonlinear models with extreme learning machines," *Neurocomputing*, vol. 102, pp. 111–124, 2013.

[24] M. Sandri and P. Zuccolotto, "Variable Selection Using Random Forests," in: S. Zani, A. Cerioli, M. Riani, and M. Vichi (eds.), *Data Analysis, Classification and the Forward Search*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, pp. 263–270, 2006.

[25] G. C. Cawley, N. L. C. Talbot, and M. Girolami, "Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation," in: B. Schölkopf, J. C. Platt, and T. Hoffmann (eds.), Advances in Neural Information Processing Systems, MIT Press, pp. 209–216, 2007.

[26] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Briefings in Bioinformatics*, vol. 9, no. 5, pp. 392–403, 2008.

[27] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[28] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in: *Proc. 18th International Conference on Machine Learning (ICML-2001)*, San Francisco, CA, USA, Morgan Kaufmann, pp. 74–81, 2001.

[29] J. M. Cadenas, M. C. Garrido, and R. Martínez, "Feature subset selection Filter–Wrapper based on low quality data," *Expert Systems with Applications*, vol. 40, pp. 6241–6252, 2013.

[30] I. S. Oh, J. S. Lee, and B. R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1424–1437, 2004.

[31] S. I. Ali and W. Shahzad, "A Feature Subset Selection Method based on Conditional Mutual Information and Ant Colony Optimization," *International Journal of Computer Applications*, vol. 60, no. 11, pp. 5–10, 2012.

[32] S. Sarafrazi and H. Nezamabadi-pour, "Facing the classification of binary problems with a GSA-SVM hybrid system," *Mathematical and Computer Modelling*, vol. 57, issues 1-2, pp. 270–278, 2013.

[33] J. Zhou, J. Liu, V. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in: *Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Beijing, China, ACM, pp. 1095–1103, 2012.

[34] S. Perkins and J. Theiler, "Online feature selection using grafting," In: *Proc. 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, USA, AAAI Press, pp. 592–599, 2003.

[35] D. Zhou, J. Huang, and B. Schölkopf, "Learning from labeled and unlabeled data on a directed graph." In: *Proc. 22nd International Conference on Machine Learning (ICML-2005)*, Bonn, Germany, ACM, pp. 1041–1048, 2005.

[36] X. Wu, K. Yu, H. Wang, and W. Ding, "Online streaming feature selection," In: *Proceedings of the 27th international conference on machine learning (ICML-2010)*, Haifa, Israel, Omnipress, pp. 1159–1166, 2010.

[37] R. Diao, M. N. Parthalain, and Q. Shen, "Dynamic feature selection with fuzzy-rough sets," in: *Proc. IEEE International Conference on Fuzzy Systems (FUZZ IEEE 2013)*, Hyderabad, India, IEEE Press, pp. 1–7, 2013.

[38] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.

[39] T. Liu, S. Liu, and Z. Chen, "An evaluation on feature selection for text clustering," in: *Proc. 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, USA, AAAI Press, pp. 488–495, 2003.

[40] J. Bins and B. A. Draper, "Feature selection from huge feature sets," in: *Proc. 8th International Conference on Computer Vision (ICCV-01)*, Vancouver, British Columbia, Canada, IEEE Computer Society, pp. 159–165, 2001.

[41] M. Muštra, M. Grgić, and K. Delač, "Breast density classification using multiple feature selection," *Automatika*, vol. 53, pp. 1289–1305, 2012.

[42] N. Dessì, E. Pascariello, and B. Pes, "A Comparative Analysis of Biomarker Selection Techniques," *BioMed Research International*, vol. 2013, article ID: 387673, DOI: http://dx.doi.org/10.1155/2013/387673

[43] H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data of glioma," *Procedia Computer Science*, vol. 23, pp. 5–14, 2013.

[44] K. Brkić, "Structural analysis of video by histogram-based description of local space-time appearance," Ph.D. dissertation, University of Zagreb, Faculty of Electrical Engineering and Computing, 2013.