sciendo

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **19**, No 1
Sofia • 2019 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2019-0001

A Review of Feature Selection and Its Methods

B. Venkatesh, J. Anuradha

SCOPE, Vellore Institute of Technology, Vellore, TN, 632014, India
E-mail: venkatesh.cse88@gmail.com

***Abstract*:** *Nowadays, being in digital era the data generated by various applications are increasing drastically both row-wise and column wise; this creates a bottleneck for analytics and also increases the burden of machine learning algorithms that work for pattern recognition. This cause of dimensionality can be handled through reduction techniques. The Dimensionality Reduction* (*DR*) *can be handled in two ways namely Feature Selection* (*FS*) *and Feature Extraction* (*FE*). *This paper focuses on a survey of feature selection methods, from this extensive survey we can conclude that most of the FS methods use static data. However, after the emergence of IoT and web-based applications, the data are generated dynamically and grow in a fast rate, so it is likely to have noisy data, it also hinders the performance of the algorithm. With the increase in the size of the data set, the scalability of the FS methods becomes jeopardized. So the existing DR algorithms do not address the issues with the dynamic data. Using FS methods not only reduces the burden of the data but also avoids overfitting of the model.*

***Keywords*:** *Dimensionality Reduction* (*DR*), *Feature Selection* (*FS*), *Feature Extraction* (*FE*).

1. Introduction

As the data increases exponentially the quality of data required for processing by Data mining, Pattern Recognition, Image processing, and other Machine Learning algorithms decrease gradually. Bellman calls this scenario "*Curse of Dimensionality*". Higher dimension data leads to the prevalence of noisy, irrelevant and redundant data. Which intern causes overfitting of the model and increases the error rate of the learning algorithm. To handle these problems "*Dimensionality Reduction*" techniques are applied, and it is the part of the preprocessing stage. So, Feature Selection (FS) and Feature Extraction (FE) are most commonly using dimensionality reduction approaches. FS is used to clean up the noisy, redundant and irrelevant data. As a result, the performance is boosted.

In FS a subset of features are selected from the original set of features based on features redundancy and relevance. Based on the relevance and redundant features, Y u and L i u [1] in 2004 have classified the feature subset as four types. They are: 1) Noisy & irrelevant; 2) Redundant & Weakly relevant; 3) Weakly relevant and

Non-redundant; 4) Strongly relevant. The feature which did not require for predicting accuracy is known as an irrelevant feature. Some of the popular approaches that fit into filter and wrapper methods are models, search strategies, feature quality measures, and feature evaluation.

Set of features are key factors for determining the hypothesis of the predicting models. The No of features and the hypothesis space are directly proportional to each other, i.e., as the number of features increases, then the hypothesis space is also increased**.** For example, if there are *M* features with the binary class label in a dataset, then it has 2^{2^M} in the search space. The hypothesis space can further be reduced by discarding redundant and irrelevant features.

The relevancy of the feature is measured based on the characteristics of the data not by its value. Statistics is such one technique which shows the relationship between the features and its importance.

The distortion of irrelevant and redundant features is not due to the presence of un-useful information; it is because the features did not have a statistical relationship with other features. Individually any feature may be irrelevant but it is relevant when joined with other features [2].

FS methods are classified into three types, based on the interaction with the learning model such as Filter, Wrapper and Embedded Methods. In the Filter method features are selected based on statistical measures. It is independent of the learning algorithm and requires less computational time. Information gain [3], chi-square test [3], Fisher score, correlation coefficient, and variance threshold are some of the statistical measures used to understand the importance of the features. The performance of the Wrapper method depends on the classifier. The best subset of features is selected based on the results of the classifier. Wrapper methods are computationally more expensive than filter methods, due to the repeated learning steps and cross-validation. However, these methods are more accurate than the filter method. Some of the examples are Recursive feature elimination [4], Sequential feature selection algorithms [5], and Genetic algorithms. The third approach is the Embedded method which uses ensemble learning and hybrid learning methods for feature selection. Since it has a collective decision, its performance is better than the other two models. Random forest is one such example. It is computationally less intensive than wrapper methods. However, this method has a drawback of specific to a learning model.

1.1. FS procedure

It is proved from the literature that feature selection can improve the performance of prediction, scalability and generalization capability of the classifier. In knowledge discovery, FS plays a fundamental role in reducing the computational complexity, storage, and cost [6].

Fig. 1 shows the various stages of FS process [7] which is explained below. Its performance depends on the decision taken at every level.

1. Search direction. A n g et al. [7] state that the first stage of the FS process is finding the search direction and the starting point. The search directions are broadly classified into three types of forward search, backward search, and random search.

The search process can start with an empty set where new features are added recursively in every iteration such phenomenon is known as forward searching. In converse to it, the backward elimination search start with a complete set of features and features are removed iteratively until the desired subset of features is reached. The other variant approach is a random searching method, which builds the feature subset by both adding and removing of the features iteratively. After determining the search direction search strategy can be applied.

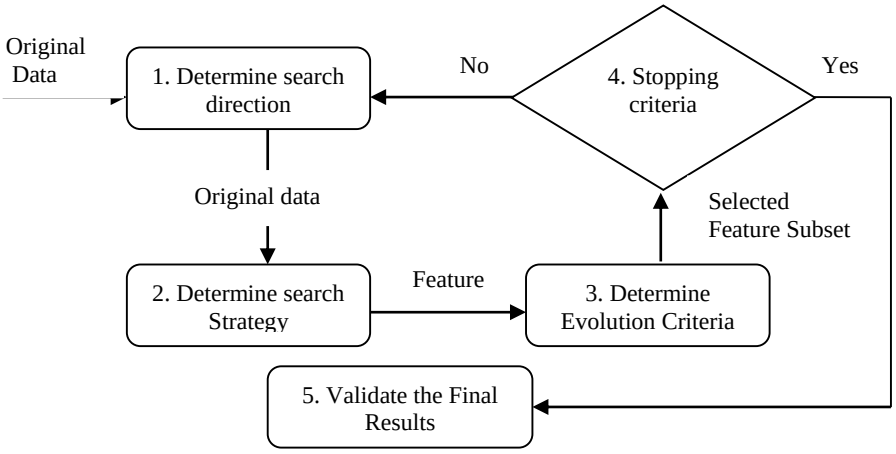

Fig. 1. Stages in FS-process

2. Determine search strategy. From the literature, we come to know that search strategies can be randomized, exponential and sequential search. Table 1 enumerates the different search strategies and their algorithms. The drawback of exponential search is that it requires 2^N combinations of feature selection for N features. It is an exhaustive search strategy, and it is an NP-hard problem [8]. To overcome this drawback researchers has introduced randomized search strategies.

In sequential search, sequentially features are added to an empty set or remove features from the complete set. Which is referred to as Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) respectively. The drawback with these methods is the features that are eliminated will not be considered for further iterations. This phenomenon is known as nesting effect. To overcome this disadvantage, F e r r i and P u d i l [9] in 1994 proposed Plus-*l*-minus-*r* ($l - r$) search method. These methods have Θ ($2M$) complexity for selecting l feature from the set of M features.

A good search strategy should obtain an optimal solution, local search ability and computational effectiveness [2]. Based on these requirements searching algorithms are further classified as optimal and suboptimal feature selection algorithms. The nesting effect of SFS and SBS algorithm are overcome by P u d i l, N o v o v i č o v á and K i t t l e r [10] in 1994 SFFS and SBFS algorithms. Different search techniques, categories, their merits, and demerits are stated in Table 2.

3. Evaluation criteria. The best features are selected based on the evaluation criteria. Based on the evaluation methods FS methods [23] are classified into Filter

method [24-26], Wrapper method [27-29], Embedded method [30], and Hybrid method [31, 32].

Table 1. Categorization of search strategies

Search strategy	Also known as	Example
Exponential search	Complete search	Exhaustive search
Sequential search	Greedy hill climbing	• Sequential Forward Selection (SFS) • Sequential Backward Selection (SBS) • Sequential Forward Floating Selection (SFFS) [10] • Sequential Backward Floating Selection (SBFS) [10] • Best first search • Beam search • Plus, *L* Take-away *r* Algorithm (PTA) [9]
Random search	Heuristic search	• Simulated annealing [11] • Random hill-climbing [12] • Genetic Algorithm (GA) [13] • Las Vegas Algorithm [14] • Tabu search [15]. • Ant Colony Optimization (ACO)[16] • Chaotic simulated annealing [17], [18] • Noisy, chaotic simulated annealing [19] • Branch-and-bound [20], [21] • Particle Swarm Optimization (PSO) [22]

Table 2. Merits and demerits of search algorithms

Search algorithm	Category	Merits	Demerits
SFS & SBS	Suboptimal FS	• Computational time is very less • Simple to implement • Robust to multi colinearity	• Redundant features are selected • Nesting effect • Fails to provide optimal results
SFFS & SFBS	Suboptimal FS	• Reduce redundant features • Overcomes Nesting effect • Provide optimal results	• NP-hard to search all subsets • A monotonic behavior of the FS
Plus-*l*-minus -*r*(*l* – *r*)	Suboptimal FS	Overcomes the nesting effect	No hypothetical method for predicting the values of *l* and *r*
Exhaustive search	Optimal FS	Reduce the computational time	• Not suitable for large dataset • NP-hard to search all subsets
Branch and bound	Optimal FS	Reduce the computational time	• Not suitable for large dataset • NP-hard to search all subsets

4. Stopping criteria. Stopping criteria specify when the FS process should stop. A good stopping criterion leads to low computational complexity in finding an optimal feature subset and also overcomes the over-fitting problem. The selection of the stopping criterion is influenced by the choices made in the previous stages. Some of the common stopping criteria are:

- Pre-defined No of features
- Pre-defined No of iterations
- Percentage (%) of advancement over two successive iteration steps
- Based on the evaluation function.

5. Validate the results. To validate the results the feature sets validation techniques are used. Cross-validation, Confusion matrix, Jaccard similarity-based measure, Rand Index are some of the validation techniques. Cross-Validation (CV) is most commonly used validation methods. The main advantage of the CV method is that it gives an unbiased error estimate. Confusion Matrix is generated for the evaluation of the classifier. Some of the classification and clustering measures commonly used are:

Classification Measures	**Clustering Measures**
Error Rate	Davies-Bouldin Index
TP Rate/ Recall / Sensitivity	Dunn Index
Specificity	F-Measure
ROC (Receiver Operating Characteristic) Curve	Jaccard index
Precision	Dice index
F-Score / F-Measure	Fowlkes-Mallows index

Further, the paper is outlined as follows: Section 2 explains how feature selection is divided based on the evaluation criteria. Section 3, elaborates how feature selection methods are applied based on the class label. Section 4 explains the applications areas of feature selection. Section 5 describes the summary and future scope of feature selection methods.

2. Feature selection based on Evaluation criteria

In this section, we are discussing the FS algorithms that depend on evaluation criteria. Based on evaluation criteria and interaction with learning algorithm feature selections are classified into three types as: 1) Filter method 2) Wrapper method, and 3) Embedded method.

2.1. Filter method

In this method, the model starts with all features and selects the best features subset based on statistical measures such as Pearson's correlation [33], Linear Discriminant Analysis (LDA), ANOVA, Chi-square [34], Wilcoxon Mann Whitney test [35], and Mutual Information (MI) [36-39]. All these statistical methods depend on the response and feature variable present in the dataset. Pearson's Correlation (PC) and Mutual Information methods are commonly using statistical methods.

2.1.1. Pearson's correlation

Correlation is a method of finding the relationship between two Quantities, for example, age and height. PC is used for detecting the linear relationship between two variables. The next equation is used to calculate the PC (ρ) between the independent variable x and dependent variable y:

$$\rho(x,y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 (y_i - \overline{y})^2}}. \quad (1)$$

Generally, the PC value lies in between [–1, 1] if the value is –1 then the variables are negatively correlated otherwise if the value is 1 then the variables are positively correlated. In case that the value is 0, then there is no correlation between the variables.

2.1.2. Mutual information

MI is another statistical method used in FS. It is the measure of how two variables (a, b) are mutually dependent. It evaluates the "measure of data" gathered about one arbitrary variable, through the other random variable. Equation 2 is used to calculate MI between two discrete random variables a and b,

$$I(A,B) = \sum_{b\in B}\sum_{a\in A} p(a,b)\log\left(\frac{p(a,b)}{p(a)p(b)}\right), \tag{2}$$

where $p(a, b)$ is the joint probability function of A and B, and $p(a)$ and $p(b)$ are the marginal probability distribution functions of A and B respectively.

For continuous random variables, the summation is replaced by a double integral as

$$I(A,B) = \int_B\int_A p(a,b)\log\left(\frac{p(a,b)}{p(a)p(b)}\right) da\, db. \tag{3}$$

In the filter method, each feature is assigned a scoring value using statistical measures. Features are organized in descending order based on the scores and assign ranking for the features. A subset of features is selected using threshold value. Filter method takes less computational time for selecting the best features. As the correlation between the independent variables is not considered while selecting the features, this leads to selection of redundant features.

Filter method uses characteristics such as information gain, consistency, dependency, correlation, and distance measures. Kira and Rendell [40] in 1992 proposed a method called *Relief* work based on instance based learning [41]. It uses Euclid distance for selecting Near-hit and Near-miss. Let X denote an instance and an instance X_i is called as Near-hit of X when it a close neighbor of X and also same class label as X. Similarly an instance X_i is called as Near-miss of X when it properly close neighbor of X and different class label as X, T denotes a relevance of threshold ranges from $0 \leq T \leq 1$. The algorithm calculates feature weight based on the average Near-hit and Near-miss. It selects the features whose feature weight is greater than T. The drawback in this algorithm is that it is applicable to two classes of classification problems and fails to discard redundant, incomplete features.

To overcome the problems with *Relief* Kononenko [42] in 1994 proposed an extension to *Relief* called *Relief-A* for addressing the incomplete data problem, *Relief-B* if at least one of two instances has unknown value for a given attribute. To address the multiclass problem, *Relief-F* is introduced. In *Relief* uses only one Near-hit/Miss for selecting the feature, whereas in *Relief-A* uses k-Nearest hits/misses and consider the average of these k-Nearest hits/misses instead of one near hit/miss. In *Relief-F* instead of finding one near miss M from a different class, the algorithm finds one near miss for each different class and averages their contribution for updating feature weights. This algorithm also fails to remove the redundant data.

Mutual Information based Feature selection was used by the Battiti [43] in 1994 to address the above-said issue. It not only finds the relevancy between the target class and individual features but also finds the relevancy between the individual features. In the selection process, the information gained by the variables helps to rule out the redundant features.

Yang and Moody [44] in 1999 proposed a method called Joint Mutual Information based approach (JMI) for the classification and visualization of non-

Gaussian data. Traditional MI has a drawback that, the method discards only a few redundant variables. For identifying the relevant feature, it calculates the JMI between the target feature and the individual features. To discard all redundant features JMI uses Conditional MI instead of normal MI.

Similarly Peng, Long and Ding [38] in 2005, proposed a heuristic algorithm called minimal-Redundancy-Maximal-Relevance (mRMR) for the removal of redundant data. It also finds most relevant features during the optimal feature subset selection, and without compromising the classification accuracy.

Based on the information theoretic (like mutual information) selection criterion Meyer and Bontempi [45] in 2006 proposed a new filter selection called Double Input Symmetrical Relevance (DISR) for supervised classifications. These methods use the variable complementary measures for finding intrinsic features and more information about the target class than that of an individual feature.

Song, Ni, and Wang [46] in 2013 proposed an algorithm for feature subset selection called a Fast clustering-based feature selection algorithm (FAST) based on graph theory. The algorithm uses the graph technique Minimum Spanning Tree (MST) for clustering the features. This FAST algorithm had effectively removed irrelevant features and redundant features by using symmetric uncertainty measure [47]. For choosing the optimal features FAST algorithm uses the cluster-based methods.

2.2. Wrapper method

According to Kohavi and John [48] in 1997 the feature subset selection in wrapper method is made as a black box, i.e., there is no knowledge about the underlying algorithm. Feature subsets are selected based on inductive algorithms. This chosen feature subset estimates the accuracy of the training model. Depending on the accuracy measured from the previous step, the method will decide whether to add or remove a feature from the selected subset. Due to this, the wrapper methods are computationally more complex.

Korfiatis et al. [49] in 2013 proposed a novel wrapper FS algorithm called LM-FM method; it comprises of two stages namely Local Maximization (LM) followed by a Floating Maximization (FM). In the LM stage best subset of features are selected from the original set of features based on credit score values between the features and the target class. Then this best subset of features is taken as input for the FM stage. In the FM stage, optimal features are selected by using floating size feature selection algorithm like Sequential Floating Forward Selection (SFFS). Korfiatis et al. combine the SVM classifier with LM-FM to show better classification.

Wrapper methods are good at classification accuracy and bad at computational efficiency. So, to overcome this problem, G. Chen and J. Chen [50] in 2015 proposed a new wrapper method namely “Cosine Similarity Measure Support Vector Machines” (CSMSVM). This technique uses the SVM classifier itself for selection of relevant features at the time of classifier construction, by including the cosine distance into SVM. This technique not only decreases the intraclass distances for the reduction of classification error rate but also it optimizes the margin in SVM. The proposed method had increased the computational efficiency to a maximum extent.

Panthong and Srivihok [51] in 2015 proposed an algorithm for wrapper FS method based on ensemble learning algorithms. In this approach, Panthong and Srivihok use three types of search strategies in the wrapper method namely SFS, SBS and Optimize selection. These methods are combined with ensemble learner. The empirical analysis of different combinations of the search strategies – Sequential and heuristic, bagging, boosting, decision tree and Naïve Bayes classifiers are used. They are: 1) FBDT (SFS + Bagging + Decision Tree), 2) BBDT (SBS + Bagging + Decision Tree), 3) OBNB (Optimize selection (evolutionary) + Bagging + Naïve Bayes), 4) OADT (Optimize selection (evolutionary) + Ada Boost + Decision Tree), and 5) OANB (Optimize selection (evolutionary) + Ada Boost+ Naïve Bayes). The study on FBDT, BBDT, OBNB, OADT, and OANB reveals that the prediction results are more accurate when combining with evolutionary algorithm – heuristic search for feature selection.

Das et al. [52] in 2017 use wrapper FS method based on harmony search. Harmony search is a meta-heuristic algorithm, which uses the concept of musical procedure for searching an idealistic harmony. In this work, instead of heuristic search, harmony search is used for subset section and is applied to identify the suitable words in native language (Bangla: Indian origin) words.

In general, the wrapper method takes more time complexity. To overcome this, Wang et al. [53] in 2017, proposed a novel approach by combing the wrapper method and filter method. This method uses the Markov blanket technique along with the wrapper-based FS for reducing the computational time. Markov blanket technique can explicitly remove the redundant feature by considering the relevance is between the features. It uses a cross-entropy based measure for this purpose. The features that are considered as redundant are conditionally independent of the target class. To reduce the no of wrapper computations the unnecessary features are identified using the filter method rather than the wrapper method.

Over past years to extract meaningful information from sensor data filter based FS method has been used. However, this approach requires more time complexity, and less occupancy estimation. Masood, Soh and Jiang [54] in 2017 have proposed a new ranking based incremental search strategy method-WRANK-ELM. This uses wrapper-ranking based feature selection, which is named after Extreme Learning Machine (ELM) classifier. The time complexity of this method is improved by adopting an incremental search strategy rather than sequential and exhaustive search. This considerably saves computational costs. The best-selected features are evaluated by using EML classifier. Experimental results of this approach outperform the results of another wrapper method.

Bermejo [55] in 2017 propose another different view on wrapper method that is applying combined wrapper methods for feature selection. This ensemble wrapper approach was tested on fish age prediction. Here Bermejo uses the ensemble techniques during the wrapper method, i.e., CV is calculated for each selected feature subsets, and the Mean CV error is calculated based on this value. Subset which has minimum CV error, that subset is considered as the best set? The collective performance of the wrapper method has shown good improvement in the accuracy.

The combination of Genetic Algorithm and Logistic Regression (GA-LR) applied by the Khammassi and Krichen [56] in 2017 was used to detect the intruders in the network. The heuristic search strategy using GA has derived the best optimal subset of features. This is evaluated by using Logistic Regression. By using this LR method, the relationship between the dependent and explanatory variables are described. This approach has capabilities of dealing with categorical data also.

2.3. Embedded method

So far, the feature selection methods that we discussed earlier use FS at the pre-processing level. The following algorithms that we are going to discuss are an embedded method. This method works in a way that the best features are selected during the learning process. The blending of feature selection during learning process has advantages of improving computational cost, classification accuracy and also avoids training the model each time when a new feature is added.

The Embedded method selects the feature subset, and the interactions of the learning algorithm were different from other feature selection methods. Filter based method learning algorithms are not used for feature selection, whereas Wrapper based method uses the learning algorithm for testing the quality of selected feature subsets. Embedded Method overcomes the computational complexity. In this method, appropriate feature selection and model learning are performed at the same time, and the features are selected during the training stage of the model. Due to this, the computational cost of this method is decidedly less compared with the wrapper method. This method avoids the training of the model each time when a new FS is explored.

Mohsenzadeh et al. [57] in 2013 proposed an algorithm called Relevant Sample-Feature Machine (RSFM) based on sparse Bayesian machine learning algorithm. The RSFM based learning model is sparse due to the adoption of Gaussian priors and Bayesian approach. RSFM is an extension of the Relevance Vector Machine (RVM) [58] algorithm; it is a sparse kernel-based learning method. In this method, the output is predicted by using the kernel function $f(x)$, i.e.,

$$f(x \mid w) = w_0 + \sum_{M=1}^{M} w_m k(x, x_n), \tag{4}$$

where $k(x, x_n)$ are kernel function and $w = (w_0, w_1, ..., w_m)^{\mathrm{T}}$, the weight vector.

Mirzaei, Mohsenzadeh and Sheikhzadeh [59] in 2017 proposed an Embedded FS method called Variational RSFM commonly referred to as VRSFM which is based on a Bayesian model of RSFM [57]. The proposed feature selection method is used for both classifications as well as regression. It defines prior Gaussian distribution on the model parameters and its hyper-parameters. For finding the hyper-parameters and posterior distributions of the parameters Mirzaei, Mohsenzadeh and Sheikhzadeh [59] employs Variational Bayesian approximation. The algorithm works well for small size dataset.

Table 3. Merits and demerits of FS method

FS method	Strengths	Gaps
Filter method	Efficient and computationally faster. Independent of the learning algorithm. Computationally faster than the Wrapper and Embedded methods. Suitable for low dimensional data	It does not consider the correlation between classifiers. It does not consider the correlation between the features. Fails to recognize the patterns properly during the learning phase
Wrapper method	It considers the correlation between the features and class labels. Also Considers the dependencies between the features. More accurate than Filter method	Computationally more complex Iteratively evaluate the selected feature subset. Some features may not be considered for evaluation when dropped at the initial stage. Searching overhead. Causes overfitting
Embedded method	Computationally efficient than the Wrapper method. More accurate than Filter and Wrapper method	Computationally costlier than the filter method. Not suitable for high dimensional data. Poor generality

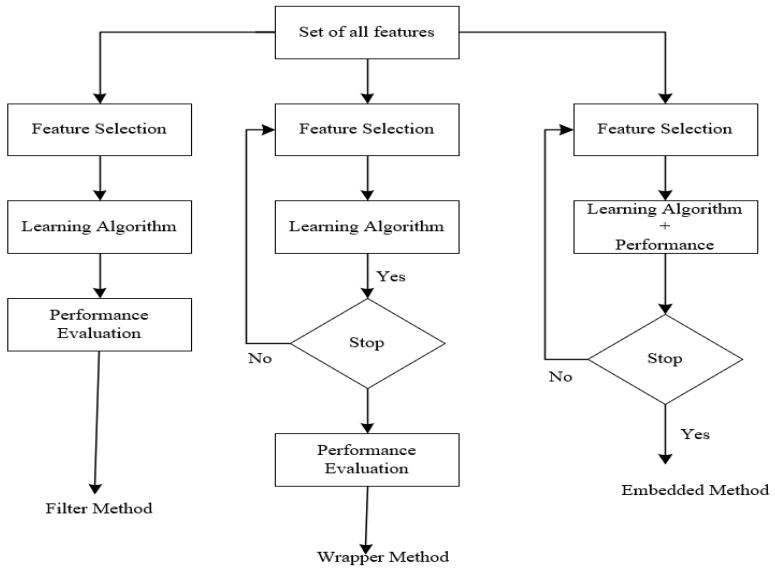

Fig. 2. Overview of Filter, Wrapper, and Embedded methods

The strengths and gaps of the FS methods are listed in Table 3. Fig. 2 represents the steps involved in the process of Filter, Wrapper, and Embedded methods.

2.4. Findings

From the literature survey above we came to know that most of the filter methods are using statistical measures for feature selection. Kira and Rendell [40] in 1992 use Relief method based on distance measure but this method fails to discard the redundant, incomplete feature. To address these problems, Kononenko [43] in

1994 extended the Relief to Relief – A, B, F for address the incomplete, unknown and multiclass problems respectively. But this method also fails to remove redundant data, so Battiti [43] in 1994 use MI-based FS to rule out the redundant features. All these methods are computationally faster but lack in the accuracy of the model. To overcome these drawbacks wrapper-based, FS are introduced.

In the wrapper method, the features are selected based on the underlying learning algorithm but it is computationally slow due to iteratively selecting for the best subset of features. Initially day sequential search strategies are used for selecting the subset of features. But it has a nesting effect to overcome this Pudil, Novovičová, and Kittler [10] in 1994 came with the idea of SFFS (Sequential Forward Floating Selection) and SBFS (Sequential Backward Floating Selection). These methods also have the drawback of the searching overhead. To overcome this Heuristic Search and optimal search strategies based on a bio-inspired algorithm for selection of optimal features with less overhead are applied. So, there is a lot of future scope in the optimization of the search strategies for better selection of the feature subsets.

There is another scope of combining the filter and wrapper to form a hybrid algorithm for better accuracy and time complexity.

3. FS methods based on Learning method

Based on the presence and absence of class labels feature selection methods Supervised FS, Unsupervised FS follow, respectively. When the dataset has both labeled and unlabelled data Semi-supervised Feature selection can be used.

3.1 Supervised FS

This approach uses the class label for selecting relevant features. Most of the time this approach causes overfitting problem due to the presence of the noisy data in the dataset. Some of the widely used supervised Feature selection methods are the Fisher score [60], Hilbert-Schmidt Independence Criterion (HSIC) [61], Fisher Criterion [62], Pearson Correlation Coefficient [63], trace ratio criterion [64] and mutual information [38].

Song et al. [61] in 2007 proposed a supervised feature selection method called BAHSIC. The dependence is estimated by using the Hilbert-Schmidt Independence Criterion (HSIC) [65], and the features are selected using backward elimination. HSIC kernel is used for measuring the dependencies. Most of the feature selection methods are applicable either for binary classification or regression but not both. The BAHSIC method has the advantage of being applied to problems of regression, binary class and multi-class classification with less computational time compared to other FS methods.

Tutkan, Ganiz and Akyokuş [66] in 2016 proposed a novel feature selection method called Meaning Based Feature Selection (MBFS) for text mining that uses Supervised and Unsupervised learning. MBFS was based on the Helmholtz principle [67] and Gestalt theorem of human perception [68], for selecting the features it uses meaning measure. Helmholtz principle from the Gestalt theory is used

for assigning a meaning score for each word in the document. For measuring the meaning score is used the next equation:

$$\text{meaning}(w, c_j) = -\frac{1}{m}\log\left[\frac{k}{m}\right] - [(m-1)\ \log N], \quad (5)$$

where w is a feature that appears k times in s dataset, m times in a document of c_j class and N is the rate of length of dataset.

Martín-Smith et al. [69] in 2017 used supervised filter method for the classification of a Brain-Computer Interface (BCI) by using Linear Discriminant Analysis (LDA) classifier. It extracts the features from ElectroEncephaloGram (EEG) signals, for analyzing the extracted signals Multi-Resolution Analysis (MRA) method had used. The proposed Filter approach had improved the formulation of multi-objective FS. For obtaining an optimal feature subset [69] had multi-objectives, they are first the method should increase the accuracy of the classifier and second the method should have overcome the overfitting problems. These had been achieved by evaluating the classifier and adjusting the parameters suitable during the training phase.

In the past Spectral Feature Selection (SFS) had been used in feature selection. But it fails to preserve either local or global structure of the data-set in the form of graph matrix. Another drawback is that it uses the original data for matrix learning every time. To overcome the problems with SFS [70] propose a novel supervised feature selections method by preserving both global and the local structure of the data set. For maximizing the objective function at a fast rate, it uses an optimization method. The graph matrix learning and the low-dimensional feature space learning are coupled as a unified framework. For preserving the global and local structure, it uses subspace-learning methods like LDA and Locality Preserving Projection (LPP) respectively, LDA uses low-rank constraint whereas LPP uses graph structure learning and for eliminating the irrelevant features, it uses $l_{2,1}$ – norm regularizer for sparse feature selection.

3.2. Graph-based unsupervised FS

In Unsupervised feature selection, the data set is unassisted by the class label, so it was the most challenging task than Supervised and Semi-supervised FS. Based on the similarity measures the redundant features are removed.

If the features are similar with one or more features then one of these features are removed, similarly, if a feature did not make any contribution to clustering, then such features are eliminated during feature selection process. It is essential for exploratory data analysis of biological data and, also useful for effectively finding unknown disease types. There are some demerits with this Unsupervised feature selection; The selected subsets did not consider the correlation between different features. Some of the well known unsupervised feature selection algorithms are Variance Score [71], Unsupervised Feature Selection using Feature Similarity measure (FSFS) [72], Laplacian Score for Feature Selection (LSFS) [73], Spectral analysis based feature selection [74], Multi-Cluster Feature Selection (MCFS) [75], and Unsupervised Discriminative Feature Selection (UDFS) [76].

In earlier unsupervised feature selection [73], He, Cai, and Niyogi 2006 uses feature ranking based techniques as fundamental criteria for feature selections. As the

feature measures are calculated independently, the relationship between them is not considered. To overcome this issue, Z. Li et al. [77] in 2012 and Y. Yang et al. [76] in 2011 have proposed a spectral-based clustering approach. In this method, the cluster structure of the data had been explored by using matrix factorization. The features are selected by using the sparsity regularization model based on learned graph Laplacian.

Bandyopadhyay et al. [78] in 2014 use dense sub-graph based on feature clustering for unsupervised feature selection. In this method first, the original feature set is represented in the form of a graph, it consists of all features being portrayed as vertices of the graph, and the edge weights are found by using the inter-feature similarity. What is computed by using mutual information? In this method the feature selection is performed in two stages, the first stage, densest sub-graph had been obtained with nonredundant features and the second stage minimizes the feature sets by using feature clustering from the graph.

X. Wang et al. [79] in 2016 propose Unsupervised Spectral Feature Selection (USFS) with l_1-norm graph. It is based on the method, Spectral Embedded Clustering [80]. For selecting the discriminative features, USFS uses l_1-norm graph and spectral clustering. By using spectral clustering, the cluster indicators are obtained from the unlabelled data sets. For cross-checking, the selected features l_1-norm graph had been imposed. It is not clear whether the manifold structure with the existing spectral feature selection method could overcome this USFS manifold structure is used for clarity.

Wen et al. [81] in 2016 proposed Unsupervised Optimal Feature Selection (UOFS) for FS. UOFS is based on $l_{2,1}$-norm regularization matrix instead of l_1-norm graph. This is because in l_1-norm graph has two phases namely 1) graph construction and 2) subspace learning before classification. But these phases are not optimal for classification. So, Wen et al. [81] use $l_{2,1}$-norm based sparse representation model, and for subspace learning, it uses $l_{2,1}$-norm regularization. The sparse representation in this method had been used for feature selection and extraction for classification.

S. Wang and H. Wang [82] in 2017 proposed a novel method for unsupervised feature selection based on low-rank approximation and structure learning. Using low-rank approximation one can provide an exact evaluation for the number of Connected components of embedded graphs in structure learning. The primary step in this method is to represent the feature selection problem as a matrix factorization with low-rank constraints by using a self-representation of a data matrix. For capturing the sparsity of the feature selection matrix, $l_{2,1}$-norm method had been used. Based on these structured learning and low-rank approximation techniques, an efficient algorithm had been implemented. There are some demerits with this method namely, how to learn the feature subsets adaptively.

Y. Liu et al. [83] in 2017 proposed a novel method for Unsupervised feature selection called Diversity-Induced Self-Representation (DISR), based on Self representation property [84] and also used an algorithm called Augmented Lagrange Method (ALM) for efficient optimization. By using diversity property, more information about the data can be captured, which helps in discarding the similar features. Due to this redundant features are significantly removed. The similarity

between the *m*-th and *n*-th features can be calculated using dot product weight as $s_{mn} = f_m^t f_n,\ m,\ n = 1, 2, \cdots,\ i$. A larger s_{mn} means *m*-th and *n*-th features are more similar. For selecting the most valuable features, it uses both diversity properties and representativeness.

Hu et al. [85] in 2017 proposed a novel method called Graph Self Representation Sparse Feature Selection (GSR-SFS) for Unsupervised Feature selection. For improving the stability of feature, the selection is achieved by integrating a subspace-learning model, (i.e., LPP) into a sparse feature level self-representation method. To achieve interpretation ability the technique uses the feature level self-representation loss function, similarly to produce stability for subspace learning it uses $l_{2,1}$-norm regularization.

Du et al. [86] in 2017 proposed Robust Unsupervised Feature Selection via Matrix Factorization (RUFSM) method for unsupervised feature selection. The data matrix is decomposed into two matrices which contain latent cluster centers and sparse representation using $l_{2,1}$-norm. High accurate discriminative feature selection is achieved by estimating the orthogonal cluster centers.

Qi et al. [87] in 2017 proposed a novel method called Regularized Matrix Factorization Feature Selection (RMFFS). Matrix factorization determines the correlation among features. For making the feature weight matrix as sparsity, it considers the absolute values of the inner product of the feature weight matrix. The combination of l_1-norm and l_2 norm is used for matrix factorization.

3.2.1. Findings

He, Cai and Niyogi [73] in 2006 use rank based for feature selections, but they do not consider the relationship between them. Li et al. [77] in 2012 and Yang et al. [76] in 2011 address these issues by using spectral based clustering approach. Later most of the researchers use graph-based learning for feature clustering. Initially, l_1-norm based graph learning was used. But using this the classifications are not optimal to overcome this $l_{2,1}$-norm graph-based methods are introduced. $l_{2,1}$ based method is gaining more popular due to its optimal feature selection.

3.3. Semi-supervised feature selection

In Semi-supervised feature selection, the learning data set contains both labeled and unlabeled data.

The graph Laplacian methods have gained the attention of many researchers working on semi-supervised based feature selection. The weighted graph is constructed for the given data for which the feature selection is applied [88].

There are mainly three stages of Graph based on Semi-Supervised Learning (GSSL). Firstly, assessing the proclivity (affinity) between a pair of samples (or) sets the users may choose kernel or similarity function. Belkin and Niyogi [89] in 2008 used a Gaussian kernel model as a similarity function, and his empirical study proves its performance. Secondly, the users have to choose an appropriate algorithm for the construction of sparse weighted subgraph from the completely weighted graph between all set of nodes. Some of the regularly used algorithms for the development of sparse subgraphs are k-Nearest Neighbour (k-NN) and Ɛ-Neighbourhood. Finally,

the user has to use a graph-based SSL algorithm for diffusing the class labels on the known node of the graph to the unknown data nodes.

Some of the graph based SSL algorithms are graph min-cut method [90], Gaussian fields and harmonic methods [91], the global and local consistency method [92], the manifold regularization [89] and the alternating graph transduction method [93].

Most of the methods in GSSL follow neighborhood methods like k-NN. From the literature, it is clear that the neighborhood approach constructed using GSSL generates irregular and imbalanced graphs for real and synthetic data. The greedy approach in adding nodes to the graph, which is based on k nearest points, is the cause for the above-said issue. To overcome the drawback of k-NN based GSSL in [93] proposed a method named maximum weight *b*-matching. In this method, each node had precisely *b* nodes in the graph.

Z h a o and L i u [94] in 2007 had proposed an algorithm called sSelect for semi-supervised learning based on spectral graph analysis. The algorithm ranks the features in the way similar to Fisher score by using a regularization framework. This algorithm selects the features one by one without considering the relationship between the features.

Generally, this graph based semi-supervised feature selection has broad applications in the area of image annotation. M a et al. [95] in 2012 Proposed an algorithm called Structural Feature Selection with Sparsity (SFSS), by using automatic image annotation. Y. Y a n g et al. [96] in 2012 used a joint framework by joining shared structure learning and graph-based learning for annotating the web images. For annotating the noisily tagged web images, T a n g et al. [97] in 2011 have proposed an algorithm called a novel k-NN sparse graph-based SSL approach.

3.3.1. Findings

From the literature, we came to know that there are mainly two drawbacks to the graph-based semi-supervised feature selection. First, these methods are not suitable for the large-scale data set, due to the presence of a large number of training datasets and also due to that they consume more time for the construction of graph like Laplacian matrix. Second, it selects the features independently without considering the correlation between the features.

4. Applications

These days there is a demand for computational power, processing capacity and storage to handle the volume of data in various fields such as massive Image processing, Microarray data, Graph Theory, Gene Selection, Network security and so on. The massive data is the major concern for the learning models. To improve the performance of the learner, it is very much essential to apply dimensionality reduction techniques to generate compact and error-free data for better results. In the following paragraphs we explain in details about each application area in details.

4.1. Hyperspectral images

In the standard image, only RGB spectral bands will be present whereas in Hyperspectral images there are several hundreds of spectral bands available. So, each pixel is used for the characterization of the objects. These hyperspectral images have been widely used in applications like remote sensing, medical imaging and so on. In hyperspectral images, the data contains rich information for different applications, but not all the measures are crucial for a particular application. Due to the presence of a large number of spectral bands, this leads to presence of redundancy between these bands. So, there is a need for feature selection method for the elimination of the redundant bands.

So, feature selection techniques are essential for selecting the relevant subset of features. Gabor wavelet transformation-based feature extraction has increased the performance of the classifiers, but in this method, too many Gabor features are extracted which makes the burden on the computation and efficiency of the technique. To overcome this in [98] a multi-task joint sparse representation framework based Gabor cube feature selection is proposed. In [99] is explained about various issues and the challenges in FS for hyperspectral image analysis.

4.2. Intrusion detection

Nowadays network-based technologies are increasing rapidly and the attacks on these techniques also increases. To overcome this problem, we should have to build a high secured Intrusion Detection System (IDS). These IDS need to deal with high dimensions of data which contain noisy, redundant and irrelevant data. This leads to a decrease in the intrusion detection rate and requires more computation time. So to achieve high detection rate FS methods are needed.

Amiri et al. [100] in 2011 propose an FS method for IDS using Mutual Information a filter method approach. Moreover, Y. Chen et al. [101] in 2006 explain about different feature selections available for IDS. From the literature, it is clear that hybrid based methods are more reliable and suitable for this application when compared to the wrapper method. However, wrapper based methods are useful when the data size is small. Whereas the filter method is used for the fast computational process, but it is less accurate.

4.3. Microarray data

Generally, microarray gene selection data consists of hundreds and thousands of features and have few rows. This becomes challenging for the learning models, so there is a need for reducing the dimensions of the data. Ang et al. [7] in 2016 clearly explain various gene selection methods for supervised, unsupervised and semi-supervised based learning models. Mandal and Mukhopadhyay [102] in 2013 proposed an improved mRMR feature selection for gene expression data. In the existing literature, most of the methods use either redundancy or relevance feature selection methods. Whereas Mandal and Mukhopadhyay, 2013 proposed a method where redundancy and relevances are considered parallely.

Interestingly, semi-supervised and unsupervised feature selection in selecting the gene features outperforms the supervised feature selection models. Other methods

like hybrid and ensemble frameworks are also producing more significant and good classification results. Few researchers are only attempted on hybrid and ensemble approaches and showed that it gives the promising result. Therefore there is more scope for improvement in this lines of selection.

Table 4. Applications of FS methods

Author & Year	Application	Algorithm	Approach
Huerta, Duval and Hao [103], 2006	Microarray data	Genetic Algorithm	Hybrid
Duval, Hao and Hernandez [104], 2009	Microarray	Genetic Algorithm and iterated local search	Embedded
Chuang, Yang and Yang [105], 2009	Microarray data	PSO + Tabu search	Wrapper
Jirapech-Umpai and Aitken [106], 2005	Microarray	Genetic Algorithm	Wrapper
Roffo and Melzi [107], (2016)	Microarrays	Eigenvector Centrality FS	Filter
Du et al. [86], 2017	Handwritten digit	RUFSM	Unsupervised FS
Peng, Long and Ding [38], 2005	Handwritten digits	mRMR	Wrapper
Oh, Lee and Suen [108], 1999	Handwriting recognition	Class dependent features	FS
Kapetanios [109], 2005	Economy	Simulated Annealing and Genetic Algorithm	Wrapper
Al-Ani [110], 2005	Texture classification	Ant Colony Optimization	Feature subset selection
Shen et al. [111], 2013	Hyperspectral image classification	Symmetrical Uncertainty (SU) and Approximate Markov Blanket (AMB)	Filter
Yao et al. [112], 2017	Image recognition	Locally Linear Embedding (LLE)	Filter
Zhang et al. [113], 2014	Spam detection	Mutation + Binary PSO	Wrapper
Ambusaidi et al. [114], 2016	Intrusion detection	Mutual Information	Filter
Alonso-Atienza et al. [115], 2014	Detection of life-threatening arrhythmias	F-score and mRMR	Filter
Roffo, Melzi and Cristani [116], 2015	Computer vision	Infinite FS	Filter
Zhang et al. [117], 2015	Alzheimer's disease	Welch's t-test	Filter
Martin-Smith et al. [69], 2017	Brain-computer interface	Linear discriminant analysis	Supervised FS
Li et al. [118], 2017	Fault detection and diagnosis	Information Greedy Feature Filter (IGFF)	Filter

In Table 4 the different FS methods and its applications are presented. The filter-based approaches are widely used methods.

5. Conclusion and future scope

As we are in the digital era every moment it generates million, billion of data. This increases the burden for processing, which in turn affect the decision making on any

application. This drags the attention of the researchers to come up with the best feature selection model that suits for any application irrespective of the constraints. So, we need to reduce the dimensions of the data by using some of the dimensionality reduction methods mentioned above.

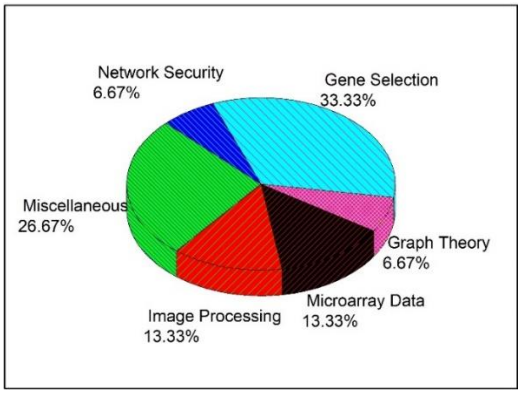

Fig. 3. Ratio of FS approaches used in different domains

It is observed from the literature, that filter-based feature selections are computationally faster when compared with the wrapper method and less accurate. Whereas in wrapper method the accuracy is more but computationally costlier. Dimension reduction provides several advantages: it results in a low dimensional model, requires less memory space, reduces the risk of overfitting, better accuracy and reduces the time complexity.

From Fig. 3 it specifies most of the researchers use Gene selection as an application area. Moreover, from Fig. 4 correlation criteria algorithm had been used by most of the researchers.

After doing a critical literature survey, it is clear that most of the experimental analysis is carried out on the static dataset. In reality, many applications generate dynamic and live data, which tends to drift the concept frequently. So, there is scope for understanding the concept and propose suitable dimensionality reduction model.

In wrapper feature selection methods, sequential search is using for the selection of feature subsets. Due to this, it increases the time complexity. To overcome this problem some of the researchers introduce a genetic algorithm-based searching methods like Ant Colony Optimization; Practical Swarm Optimization are some of the commonly used methods. By using the optimization methods, there is a significant improvement in the feature subset selection. So most of the future research works are carried out by using this methods only. There is more future scope in this area.

Another area of feature scope is using hybrid methods by combining filter and wrapper method for better performance of the reduction techniques. Here in this hybrid methods the researcher is coming with filter and wrapper method to improve the performance of the classification algorithms. Using this approach also has a more future scope.

Finally, another area of future scope for dimensionality reduction is using of graph based feature selection for unsupervised feature selction.

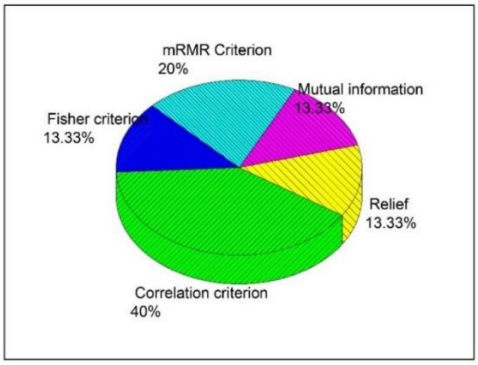

Fig. 4. Comparison of FS methods

References

1. Yu, L., H. Liu. Efficient Feature Selection via Analysis of Relevance and Redundancy. – J. Mach. Learn. Res., Vol. **5**, 2004, No Oct, pp. 1205-1224.
2. Gheyas, I. A., L. S. Smith. Feature Subset Selection in Large Dimensionality Domains. – Pattern Recognit., Vol. **43**, January 2010, No 1, pp. 5-13.
3. Yang, Y., J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. – In: Proc. of 14th International Conference on Machine Learning, ICML'97, 1997, pp. 412-420.
4. Yan, K., D. Zhang. Feature Selection and Analysis on Correlated Gas Sensor Data with Recursive Feature Elimination. – Sensors Actuators, B Chem., Vol. **212**, Jun 2015, pp. 353-363.
5. Jain, A., D. Zongker. Feature Selection: Evaluation, Application, and Small Sample Performance. – IEEE Trans. Pattern Anal. Mach. Intell., Vol. **19**, 1997, No 2, pp. 153-158.
6. Gutkin, M., R. Shamir, G. Dror. SlimPLS: A Method for Feature Selection in Gene Expression-Based Disease Classification. – PLoS One, Vol. **4**, July 2009, No 7, p. e6416.
7. Ang, J. C., A. Mirzal, H. Haron, H. N. A. Hamed. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. – IEEE/ACM Trans. Comput. Biol. Bioinforma., Vol. **13**, September 2016, No 5, pp. 971-989.
8. Bins, J., B. A. Draper. Feature Selection from Huge Feature Sets. – In: Proc. of IEEE Int. Conf. Comput. Vis., Vol. **2**, 2001, pp. 159-165.
9. Ferri, F., P. Pudil. Comparative Study of Techniques for Large-Scale Feature Selection. – Pattern Recognit. Pract. IV, Vol. **1994**, 1994, pp. 403-413.
10. Pudil, P., J. Novovičová, J. Kittler. Floating Search Methods in Feature Selection. – Pattern Recognit. Lett., Vol. **15**, November 1994, No 11, pp. 1119-1125.
11. Doak, J. An Evaluation of Feature Selection Methodsand Their Application to Computer Security. CSE-92-18, 1992. 82 p.
12. Skalak, D. B. Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms. – In: Proc. of 11th International Conference on Machine Learning, 1994, pp. 293-301.
13. Goldberg, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning. Boston, MA, 1989. – Read. Addison-Wesley, 1989.
14. Brassard, P., Gilles, Bratley. Fundamentals of Algorithmics. Englewood Cliffs, NJ, Prentice Hall, 1996.
15. Glover, F. Future Paths for Integer Programming and Links to Artificial Intelligence. – Comput. Oper. Res., Vol. **13**, January 1986, No 5, pp. 533-549.
16. Li, B., L. Wang, W. Song. Ant Colony Optimization for the Traveling Salesman Problem Based on Ants with Memory. – In: Proc. of 4th International Conference on Natural Computation, 2008, pp. 496-501.

17. N o z a w a, H. A Neural Network Model as a Globally Coupled Map and Applications Based on Chaos. Chaos an Interdiscip. – J. Nonlinear Sci., Vol. **2**, July 1992, No 3, pp. 377-386.
18. L u o n a n, C., K. A i h a r a. Chaotic Simulated Annealing by a Neural Network Model with Transient Chaos. – Neural Networks, Vol. **8**, 1995, No 6, pp. 915-930.
19. W a n g, L., S. L i, F. T i a n, X. F u. A Noisy Chaotic Neural Network for Solving Combinatorial Optimization Problems: Stochastic Chaotic Simulated Annealing. – IEEE Trans. Syst. Man, Cybern. Part B Cybern., Vol. **34**, 2004, No 5, pp. 2119-2125.
20. N a r e n d r a, P. M., K. F u k u n a g a. A Branch and Bound Algorithm for Feature Subset Selection. – IEEE Trans. Comput., Vol. **C-26**, 1977, No 9, pp. 917-922.
21. L a n d, A., A. D o i g. An Automatic Method of Solving Discrete Programming Problems. – Econometrika, Vol. **28**, 1960, No 3, pp. 497-520.
22. P o l i, R., J. K e n n e d y, T. B l a c k w e l l. Particle Swarm Optimization. – Swarm Intell., Vol. **1**, October 2007, No 1, pp. 33-57.
23. D a s h, M., H. L i u. Feature Selection for Classification. – Intell. Data Anal., Vol. **1**, January 1997, No 1-4, pp. 131-156.
24. F a y y a d, M. U., K. B. I r a n i.The Attribute Selection Problem in Decision Tree Generation. – Aaai-92, 1992, pp. 104-110.
25. L i u, H., R. S e t i o n o. A Probabilistic Approach to Feature Selection – A Filter Solution. – In: Proc. of 13th International Conference on Machine Learning, 1996, pp. 319-327.
26. S i e d l e c k i, W., J. S k l a n s k y. On Automatic Feature Selection. – Int. J. Pattern Recognit. Artif. Intell., Vol. **02**, Jun 1988, No 02, pp. 197-220.
27. D y, J. G., C. E. B r o d l e y. Feature Subset Selection and Order Identification for Unsupervised Learning. – In: Proc. of 17th Int. Conf. Mach. Learn ICML'00, 2000, pp. 247-254.
28. J o h n, G. H., R. K o h a v i, K. P f l e g e r. Irrelevant Features and the Subset Selection Problem. – In: Machine Learning Proceedings 1994, 1994, pp. 121-129.
29. C a r u a n a, R., D. F r e i t a g. Greedy Attribute Selection. – In: Proc. Elev. Int. Conf. Mach. Learn., Vol. **48**, 1994, pp. 28-36.
30. A s i r, D., S. A p p a v u, E. J e b a m a l a r. Literature Review on Feature Selection Methods for High-Dimensional Data. – Int. J. Comput. Appl., Vol. **136**, February 2016, No 1, pp. 9-17.
31. D a s, S. Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. – Engineering, 2001, pp. 74-81.
32. T a l a v e r a, L. Feature Selection as a Preprocessing Step for Hierarchical Clustering. – In: Proc. of 25th Int. Conf. Mach. Learn., 1999, pp. 389-397.
33. B i e s i a d a, J., W. D u c h. Feature Selection for High-Dimensional Data – A Pearson Redundancy Based Filter. – In Advances in Soft Computing, Vol. **45**, Springer, Berlin, Heidelberg, 2007, pp. 242-249.
34. J i n, X., A. X u, R. B i e, P. G u o. Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles. – In: Proc. of 2006 International Conference on Data Mining for Biomedical Applications, Springer-Verlag, 2006, pp. 106-115.
35. L i a o, C., S. L i, Z. L u o. Gene Selection Using Wilcoxon Rank Sum Test and Support Vector Machine for Cancer Classification. – Comput. Intell. Secur., Vol. **4456**, 2007, pp. 57-66.
36. V i n h, L. T., N. D. T h a n g, Y.-K. L e e. An Improved Maximum Relevance and Minimum Redundancy Feature Selection Algorithm Based on Normalized Mutual Information. – In: Proc. of 10th IEEE/IPSJ International Symposium on Applications and the Internet, 2010, pp. 395-398.
37. E s t e v e z, P. A., M. T e s m e r, C. A. P e r e z, J. M. Z u r a d a. Normalized Mutual Information Feature Selection. – IEEE Trans. Neural Networks, Vol. **20**, February 2009, No 2, pp. 189-201.
38. P e n g, H., F. L o n g, C. D i n g. Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. – IEEE Trans. Pattern Anal. Mach. Intell., Vol. **27**, August 2005, No 8, pp. 1226-1238.
39. K w a k, N., C h o n g-H o C h o i. Input Feature Selection by Mutual Information Based on Parzen Window. – IEEE Trans. Pattern Anal. Mach. Intell., Vol. **24**, December 2002, No 12, pp. 1667-1671.

40. K i r a, K., L. R e n d e l l. A Practical Approach to Feature Selection. – In: Proc. of 9th Int'l Workshop on Machine Learning, 1992, pp. 249-256.
41. A h a, D. W., D. K i b l e r, M. K. A l b e r t. Instance-Based Learning Algorithms. – Mach. Learn., Vol. **6**, January 1991, No 1, pp. 37-66.
42. K o n o n e n k o, I. Estimating Attributes: Analysis and Extensions of RELIEF. Berlin, Heidelberg, Springer, 1994, pp. 171-182.
43. B a t t i t i, R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. – IEEE Trans. Neural Networks, Vol. **5**, July 1994, No 4, pp. 537-550.
44. Y a n g, H. H., J. M o o d y. Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. – In: In Advances in Neural Information Processing Systems, 1999, pp. 687-693.
45. M e y e r, P. E., G. B o n t e m p i. On the Use of Variable Complementarity for Feature Selection in Cancer Classification. – In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. **3907**. LNCS, Springer, Berlin, Heidelberg, 2006, pp. 91-102.
46. S o n g, Q., J. N i, G. W a n g. A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. – IEEE Trans. Knowl. Data Eng., Vol. **25**, January 2013, No 1, pp. 1-14.
47. P r e s s, W. H., S. A. T e u k o l s k y, W. T. V e t t e r l i n g, B. P. F l a n n e r y. – Numerical Recipes. 2nd Ed. Cambridge, Cambridge University Press, 1989.
48. K o h a v i, R., G. H. J o h n. Wrappers for Feature Subset Selection. – Artif. Intell., Vol. **97**, December 1997, No 1-2, pp. 273-324.
49. K o r f i a t i s,V. C., P. A. A s v e s t a s, K. K. D e l i b a s i s, G. K. M a t s o p o u l o s. A Classification System Based on a New Wrapper Feature Selection Algorithm for the Diagnosis of Primary and Secondary Polycythemia. – Comput. Biol. Med., Vol. **43**, December 2013, No 12, pp. 2118-2126.
50. C h e n, G., J. C h e n. A Novel Wrapper Method for Feature Selection and Its Applications. – Neurocomputing, Vol. **159**, July 2015, No 1, pp. 219-226.
51. P a n t h o n g, R., A. S r i v i h o k. Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm. – Procedia Comput. Sci., Vol. **72**, 2015, pp. 162-169.
52. D a s, S., P. K. S i n g h, S. B h o w m i k, R. S a r k a r, M. N a s i p u r i. A Harmony Search Based Wrapper Feature Selection Method for Holistic Bangla Word Recognition. – Procedia Comput. Sci., Vol. **89**, July 2017, pp. 395-403.
53. W a n g, A., N. A n, J. Y a n g, G. C h e n, L. L i, G. A l t e r o v i t z. Wrapper-Based Gene Selection with Markov Blanket. – Comput. Biol. Med., Vol. **81**, 2017, pp. 11-23.
54. M a s o o d, M. K., Y. C. S o h, C. J i a n g. Occupancy Estimation from Environmental Parameters Using Wrapper and Hybrid Feature Selection. – Appl. Soft Comput. J., Vol. **60**, November 2017, pp. 482-494.
55. B e r m e j o, S. Ensembles of Wrappers for Automated Feature Selection in Fish Age Classification. – Comput. Electron. Agric., Vol. **134**, March 2017, pp. 27-32.
56. K h a m m a s s i, C., S. K r i c h e n. A GA-LR Wrapper Approach for Feature Selection in Network Intrusion Detection. – Comput. Secur., Vol. **70**, September 2017, pp. 255-277.
57. M o h s e n z a d e h, Y., H. S h e i k h z a d e h, A. M. R e z a, N. B a t h a e e, M. M. K a l a y e h. The Relevance Sample-Feature Machine: A Sparse Bayesian Learning Approach to Joint Feature-Sample Selection. – IEEE Trans. Cybern., Vol. **43**, 2013, No 6, pp. 2241-2254.
58. T i p p i n g, M. M. Sparse Bayesian Learning and the Relevance Vector Machine. – J. Mach. Learn. Res., Vol. **1**, 2001, pp. 211-245.
59. M i r z a e i, A., Y. M o h s e n z a d e h, H. S h e i k h z a d e h. Variational Relevant Sample-Feature Machine: A Fully Bayesian Approach for Embedded Feature Selection. – Neurocomputing, Vol. **241**, 2017, pp. 181-190.
60. G u, Q., Z. L i, J. H a n. Generalized Fisher Score for Feature Selection. February 2012.
61. S o n g, L., A. S m o l a, A. G r e t t o n, K. M. B o r g w a r d t, J. B e d o. Supervised Feature Selection via Dependence Estimation. – In: Proc. of 24th International Conference on Machine Learning (ICML'07), 2007, pp. 823-830.

62. L o o g, M., R. P. W. D u i n, R. H a e b-U m b a c h. Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria. – IEEE Trans. Pattern Anal. Mach. Intell., Vol. **23**, July 2001, No 7, pp. 762-766.
63. R o d g e r s, J. L., W. A. N i c e w a n d e r. Thirteen Ways to Look at the Correlation Coefficient. – Am. Stat., Vol. **42**, February 1988, No 1, p. 59.
64. N i e, F., F. N i e, S. X i a n g, Y. J i a, C. Z h a n g, S. Y a n. Trace Ratio Criterion for Feature Selection. – AAAI, 2008, pp. 671-676.
65. G r e t t o n, A., O. B o u s q u e t, A. S m o l a, B. S c h ö l k o p f. Measuring Statistical Dependence with Hilbert-Schmidt Norms. – Springer, 2005, pp. 63-78.
66. T u t k a n, M., M. C. G a n i z, S. A k y o k u ş. Helmholtz Principle Based Supervised and Unsupervised Feature Selection Methods for Text Mining. – Inf. Process. Manag., Vol. **52**, September 2016, No 5, pp. 885-910.
67. B a l i n s k y, A., H. B a l i n s k y, S. S i m s k e. On the Helmholtz Principle for Data Mining. – Hewlett-Packard Dev. Company, LP, 2011.
68. D e s o l n e u x, A., L. M o i s a n, J.-M. M o r e l. From Gestal Theory to Image Analysis: A Probabilistic Approach. 2008.
69. M a r t í n-S m i t h, P., J. O r t e g a, J. A s e n s i o-C u b e r o, J. Q. G a n, A. O r t i z. A Supervised Filter Method for Multi-Objective Feature Selection in EEG Classification Based on Multi-Resolution Analysis for BCI. – Neurocomputing, Vol. **250**, August 2017, pp. 45-56.
70. Z h u, Y., X. Z h a n g, R. H u, G. W e n. Adaptive Structure Learning for Low-Rank Supervised Feature Selection. – Pattern Recognition Letters, North-Holland, 16 August 2017.
71. B i s h o p, C. M. Neural Networks for Pattern Recognition. Clarendon Press, 1995.
72. M i t r a, P., C. A. M u r t h y, S. K. P a l. Unsupervised Feature Selection Using Feature Similarity. – IEEE Trans. Pattern Anal. Mach. Intell., Vol. **24**, March 2002, No 3, pp. 301-312.
73. H e, X., D. C a i, P. N i y o g i. Laplacian Score for Feature Selection. 2006, pp. 507-514.
74. Z h a o Z., H. L i u. Spectral Feature Selection for Supervised and Unsupervised Learning. – In: Proc. of 24th International Conference on Machine Learning (ICML'07), 2007, pp. 1151-1157.
75. C a i, D., C. Z h a n g, X. H e. Unsupervised Feature Selection for Multi-Cluster Data. – In: Proc. of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'10, 2010, p. 333.
76. Y a n g, Y., H. T. S h e n, Z. M a, Z. H u a n g, X. Z h o u. l 2,1 -Norm Regularized Discriminative Feature Selection for Unsupervised Learning. – In: Proc. of 22nd Int. Jt. Conf. Artif. Intell., Vol. **2**, 2011, pp. 1589-1594.
77. L i, Z., Y. Y a n g, J. L i u, X. Z h o u, H. L u. Unsupervised Feature Selection Using Nonnegative Spectral Analysis. – In: Proc. of 26th AAAI Conference on Artificial Intelligence, AAAI Press, 2012, pp. 1026-1032.
78. B a n d y o p a d h y a y, S., T. B h a d r a, P. M i t r a, U. M a u l i k. Integration of Dense Subgraph Finding with Feature Clustering for Unsupervised Feature Selection. – Pattern Recognit. Lett., Vol. **40**, April 2014, No 1, pp. 104-112.
79. W a n g, X., X. Z h a n g, Z. Z e n g, Q. W u, J. Z h a n g. Unsupervised Spectral Feature Selection with l1-Norm Graph. – Neurocomputing, Vol. **200**, August 2016, pp. 47-54.
80. N i e, F., Z. Z e n g, I. W. T s a n g, D. X u, C. Z h a n g. Spectral Embedded Clustering: A Framework for In-Sample and Out-of-Sample Spectral Clustering. – IEEE Trans. Neural Networks, Vol. **22**, November 2011, No 11, pp. 1796-1808.
81. W e n, J., Z. L a i, Y. Z h a n, J. C u i. The L2,1-Norm-Based Unsupervised Optimal Feature Selection with Applications to Action Recognition. – Pattern Recognit., Vol. **60**, December 2016, pp. 515-530.
82. W a n g, S., H. W a n g. Unsupervised Feature Selection via Low-Rank Approximation and Structure Learning. – Knowledge-Based Syst., Vol. **124**, May 2017, pp. 70-79.
83. L i u, Y., K. L i u, C. Z h a n g, J. W a n g, X. W a n g. Unsupervised Feature Selection via Diversity-Induced Self-Representation. – Neurocomputing, Vol. **219**, January 2017, pp. 350-363.
84. Z h u, P., W. Z u o, L. Z h a n g, Q. H u, S. C. K. S h i u. Unsupervised Feature Selection by Regularized Self-Representation. – Pattern Recognit., Vol. **48**, February 2015, No 2, pp. 438-446.

85. H u, R. et al. Graph Self-Representation Method for Unsupervised Feature Selection. – Neurocomputing, Vol. **220**, January 2017, pp. 130-137.
86. D u, S., Y. M a, S. L i, Y. M a. Robust Unsupervised Feature Selection via Matrix Factorization. – Neurocomputing, Vol. **241**, Jun 2017, pp. 115-127.
87. Q i, M., T. W a n g, F. L i u, B. Z h a n g, J. W a n g, Y. Y i. Unsupervised Feature Selection by Regularized Matrix Factorization. – Neurocomputing, Vol. **273**, 17 January 2017, Elsevier, pp. 593-610.
88. Z h u, X. Semi-Supervised Learning Literature Survey Contents. Learning, 2006.
89. B e l k i n, M., P. N i y o g i. Towards a Theoretical Foundation for Laplacian-Based Manifold Methods. – J. Comput. Syst. Sci., Vol. **74**, December 2008, No 8, pp. 1289-1308.
90. B l u m, A., S. C h a w l a. Learning from Labeled and Unlabeled Data Using Graph Mincuts. – In: ICML'01, 2001.
91. Z h u,, X., X. Z h u, Z. G h a h r a m a n i, J. L a f f e r t y. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. – In: ICML'03, 2003, pp. 912-919.
92. Z h o u, D., O. B o u s q u e t, T. N. L a l, J. W e s t o n, B. S c h ö l k o p f. Learning with Local and Global Consistency. – In: NIPS'03, 2003, pp. 321-328.
93. W a n g, J., T. J e b a r a, S.-F. C h a n g. Graph Transduction via Alternating Minimization. – In: Proc. of 25th International Conference on Machine Learning (ICML'08), 2008, pp. 1144-1151.
94. Z h a o, Z., H. L i u. Semi-Supervised Feature Selection via Spectral Analysis. – In: Proc. of 2007 SIAM International Conference on Data Mining, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2007, pp. 641-646.
95. M a, Z., F. N i e, Y. Y a n g, J. R. R. U i j l i n g s, N. S e b e, A. G. H a u p t m a n n. Discriminating Joint Feature Analysis for Multimedia Data Understanding. – IEEE Trans. Multimed., Vol. **14**, December 2012, No 6, pp. 1662-1672.
96. Y a n g, Y., F. W u, F. N i e, H. T. S h e n, Y. Z h u a n g, A. G. H a u p t m a n n. Web and Personal Image Annotation by Mining Label Correlation With Relaxed Visual Graph Embedding. – IEEE Trans. Image Process., Vol. **21**, March 2012, No 3, pp. 1339-1351.
97. T a n g, J., R. H o n g, S. Y a n, T.-S. C h u a, G.-J. Q i, R. J a i n. Image Annotation by kNN-Sparse Graph-Based Label Propagation over Noisily Tagged Web Images. – ACM Trans. Intell. Syst. Technol., Vol. **2**, February 2011, No 2, pp. 1-15.
98. J i a, S., Y. X i e, L. S h e n, L. D e n g. Hyperspectral Image Classification Using Fisher Criterion-Based Gabor Cube Selection and Multi-Task Joint Sparse Representation. – In: Proc. of 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS'15), 2015, pp. 1-4.
99. J i a, X., B.-C. K u o, M. M. C r a w f o r d. Feature Mining for Hyperspectral Image Classification. – Proc. IEEE, Vol. **101**, March 2013, No 3, pp. 676-697.
100. A m i r i, F., M. R e z a e i Y o u s e f i, C. L u c a s, A. S h a k e r y, N. Y a z d a n i. Mutual Information-Based Feature Selection for Intrusion Detection Systems. – J. Netw. Comput. Appl., Vol. **34**, July 2011, No 4, pp. 1184-1199.
101. C h e n, Y., Y. L i, X.-Q. C h e n g, L. G u o. Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System. – Inf. Secur. Cryptol., Vol. **4318**, November 2006, pp. 153-167.
102. M a n d a l, M., A. M u k h o p a d h y a y. An Improved Minimum Redundancy Maximum Relevance Approach for Feature Selection in Gene Expression Data. – Procedia Technol., Vol. **10**, January 2013, pp. 20-27.
103. H u e r t a, E. B., B. D u v a l, J.-K. H a o. A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data. Berlin, Heidelberg, Springer, 2006, pp. 34-44.
104. D u v a l, B., J.-K. H a o, J. C. H e r n a n d e z H e r n a n d e z. A Memetic Algorithm for Gene Selection and Molecular Classification of Cancer. – In: Proc. of 11th Annual Conference on Genetic and Evolutionary Computation (GECCO'09), 2009, p. 201.
105. C h u a n g, L.-Y., C.-H. Y a n g, C.-H. Y a n g. Tabu Search and Binary Particle Swarm Optimization for Feature Selection Using Microarray Data. – J. Comput. Biol., Vol. **16**, December 2009, No 12, pp. 1689-1703.
106. J i r a p e c h-U m p a i, T., S. A i t k e n. Feature Selection and Classification for Microarray Data Analysis: Evolutionary Methods for Identifying Predictive Genes. – BMC Bioinformatics, Vol. **6**, Jun 2005, No 1, p. 148.

107. R o f f o, G., S. M e l z i. Feature Selection via Eigenvector Centrality, December 2016. **pdfs.semanticscholar.org**
108. O h, I l-S e o k, J i n-S e o n L e e, C. Y. S u e n. Analysis of Class Separation and Combination of Class-Dependent Features for Handwriting Recognition. – IEEE Trans. Pattern Anal. Mach. Intell., Vol. **21**, 1999, No 10, pp. 1089-1094.
109. K a p e t a n i o s, G. Variable Selection Using Non-Standard Optimisation of Information Criteria, – Work. Pap. Queen Hapy University of London, No 533, 2005.
110. A l-A n i, A. Feature Subset Selection Using Ant Colony Optimization. – Int. J. Comput. Intell., Vol. **2**, 2005, No 1, pp. 53-58.
111. S h e n, L., Z. Z h u, S. J i a, J. Z h u, Y. S u n. Discriminative Gabor Feature Selection for Hyperspectral Image Classification. – IEEE Geosci. Remote Sens. Lett., Vol. **10**, January 2013, No 1, pp. 29-33.
112. Y a o, C., Y.-F. L i u, B. J i a n g, J. H a n, J. H a n. LLE Score: A New Filter-Based Unsupervised Feature Selection Method Based on Nonlinear Manifold Embedding and Its Application to Image Recognition. – IEEE Trans. Image Process., Vol. **26**, November 2017, No 11, pp. 5257-5269.
113. Z h a n g, Y., S. W a n g, P. P h i l l i p s, G. J i. Binary PSO with Mutation Operator for Feature Selection Using Decision Tree Applied to Spam Detection. – Knowledge-Based Syst., Vol. **64**, July 2014, pp. 22-31.
114. A m b u s a i d i, M. A., X. H e, P. N a n d a, Z. T a n. Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm. – IEEE Trans. Comput., Vol. **65**, October 2016, No 10, pp. 2986-2998.
115. A l o n s o-A t i e n z a, F., et al. Detection of Life-Threatening Arrhythmias Using Feature Selection and Support Vector Machines. – IEEE Trans. Biomed. Eng., Vol. **61**, 2014, No 3, pp. 832-40.
116. R o f f o, G., S. M e l z i, M. C r i s t a n i. Infinite Feature Selection. – In: 2015 IEEE International Conference on Computer Vision (ICCV'15), 2015, pp. 4202-4210.
117. Z h a n g, Y., et al. Detection of Subjects and Brain Regions Related to Alzheimer's Disease Using 3D MRI Scans Based on Eigenbrain and Machine Learning. – Front. Comput. Neurosci., Vol. **9**, Jun 2015, p. 66.
118. L i, D., Y. Z h o u, G. H u, C. J. S p a n o s. Optimal Sensor Configuration and Feature Selection for AHU Fault Detection and Diagnosis. – IEEE Trans. Ind. Informatics, Vol. **13**, Jun 2017, No 3, pp. 1369-1380.

Received: 02.02.2018; Second Version: 18.10.2018; Accepted: 07.02.2019