Review of feature selection approaches based on grouping of features

Cihan Kuzudisli [Corresp., 1], **Bahjat Qaqish** [2], **Burcu Bakir-Gungor** [3], **Malik Yousef** [Corresp. 4, 5]

[1] Department of Computer Engineering, Hasan Kalyoncu University, Gaziantep, Turkey

[2] Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, Chapell Hill, United States

[3] Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey

[4] Department of Information Systems, Zefat Academic College, Zefat, Israel

[5] Galilee Digital Health Research Center, Zefat Academic College, Zefat, Israel

Corresponding Authors: Cihan Kuzudisli, Malik Yousef
Email address: cihan.kuzudisli@hku.edu.tr, malik.yousef@gmail.com

With the rapid development in technology, large amounts of high-dimensional data have been generated. This high dimensionality including redundancy and irrelevancy poses a great challenge in data analysis and decision making. Feature selection (FS) is an effective way to reduce dimensionality by eliminating redundant and irrelevant data. Most traditional feature selection approaches consider all the features in order to score and rank to be able to perform feature selections either by eliminating lower ranked features or considering highly ranked features for training the machine learning classifier. In this review, we discuss an emerging approach to feature selection that is based on first grouping features, then scoring groups of features rather than scoring the full set. Despite the presence of reviews on clustering and FS algorithms, to the best of our knowledge, this is the first review focusing on FS techniques based on grouping. The main idea behind FS through grouping is to generate clusters of similar features with dissimilarity between clusters, then select representative features from each cluster. Approaches under supervised, unsupervised and integrative frameworks are explored. We hope this work's findings can guide effective design of new FS approaches using feature grouping.

Review of Feature Selection Approaches Based on Grouping of Features

Cihan Kuzudisli[1], Bahjat F. Qaqish[2], Burcu Bakir Gungor[3], Malik Yousef[4,5]

[1] Department of Computer Engineering, Hasan Kalyoncu University, Gaziantep, Turkey
[2] Department of Biostatistics, University of North Carolina at Chapel Hill, NC, Chapell Hill, USA
[3] Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey
[4] Department of Information Systems, Zefat Academic College, Zefat, Israel
[5] Galilee Digital Health Research Center, Zefat Academic College, Zefat, Israel

Corresponding Author:
Cihan Kuzudisli[1]
Gaziantep, 27010, Turkey
Email address: cihan.kuzudisli@hku.edu.tr
Malik Yousef[2]
Zefat, 13206, Israel
Email address: malik.yousef@gmail.com

Abstract

With the rapid development in technology, large amounts of high-dimensional data have been generated. This high dimensionality including redundancy and irrelevancy poses a great challenge in data analysis and decision making. Feature selection (FS) is an effective way to reduce dimensionality by eliminating redundant and irrelevant data. Most traditional feature selection approaches consider all the features in order to score and rank to be able to perform feature selections either by eliminating lower ranked features or considering highly ranked features for training the machine learning classifier. In this review, we discuss an emerging approach to feature selection that is based on first grouping features, then scoring groups of features rather than scoring the full set. Despite the presence of reviews on clustering and FS algorithms, to the best of our knowledge, this is the first review focusing on FS techniques based on grouping. The main idea behind FS through grouping is to generate clusters of similar features with dissimilarity between clusters, then select representative features from each cluster. Approaches under supervised, unsupervised and integrative frameworks are explored. We hope this work's findings can guide effective design of new FS approaches using feature grouping.

Introduction

In the current digital era, the data produced in fields such as image processing, pattern recognition, machine learning and network communication grow exponentially in terms of dimension and size. Due to this high-dimensionality, the search space is widening and extraction of valuable knowledge from the data becomes a challenging task [1]. Also, utilizing all features in a dataset is unlikely to develop a predictive model with a high accuracy. Existence of irrelevant and redundant features may weaken the generalizability of the model and decrease the overall precision of a classifier [2]. Hence, it's desired to have reduced input variables to lower computational cost of the model construction and improve the performance of the model. As such, feature selection (FS) becomes an inevitable step for domain experts and data analysts.

FS is the process of selecting the minimally sized feature subset from the original set that is optimal for the target concept. It plays a crucial role for removal of irrelevant and redundant features while keeping relevant and non-redundant ones. Irrelevant features do not alter the target concept in any way and redundant features do not contribute to the target concept [3]. These features may contain a considerable amount of noise or can be deceptive which results in significant computational overhead and poor predictor performance. Contrary to other dimensionality reduction techniques, FS preserves the semantics of the data due to no distortion in the original representation of features and hence provides interpretation of data for data scientists. Additionally, reduction in dimension by FS prevents model overfitting which leads to undesired validation results.

Although various FS techniques have been developed, traditional approaches to FS neglect structures of features during the selection process. Another issue is the acquisition or elimination of the features on an individual basis, which ignores dependency between them. Because of these reasons, correlation between features may not be detected efficiently resulting in irrelevant or redundant features in the final subset. Some studies clustered samples (observations) for improving classification performance but were not concerned with feature reduction at all.

On the other hand, feature selection based on clustering, that is, feature grouping (clustering) is an effective technique for reducing feature redundancy and enhancing classifier learning. By grouping features, the search dimension is substantially reduced. Moreover, it can reduce estimator variance [4], improve stability, and reinforce generalization capability of the model. Although there are reviews on clustering methods [5] and feature selection techniques [1], [6], to the best of our knowledge, this is the first paper making a literature review on approaches for feature selection based on grouping. Hereafter, grouping and clustering terms will be used interchangeably. In this procedure, clustering process is generally the initial step and performed to have maximal intra-class similarity (similarity in between the objects of the same cluster) and minimal inter-class similarity (i.e., objects in a cluster are more similar to those in another one) between features. These feature groups can be created by K-Means , fuzzy c-mean (FCM), hierarchical clustering , graph theory and even more. After the acquisition of these clusters, features within each cluster are scored and selected by different metrics or techniques.
The remainder of this paper is organized as follows: We will firstly give a concise overview of different feature selection methods in Section 2. In Section 3, we will present different works carried out in FS using feature grouping following the summary of traditional approaches. Then,

in Section 4, we will review different studies benefited from Recursive Cluster Elimination based on Support Vector Machine (SVM-RCE). Next, we will address, in Section 5, feature selection techniques involving both feature grouping and incorporating domain knowledge. Lastly, in Section 6, we conclude our review with further discussions and future directions.

Rationale of the review and intended audience

Nowadays, high throughput technologies output high dimensional data, which makes data acquisition and data analysis a challenging issue. Existence of irrelevant and redundant features makes it hard to infer meaningful conclusions from data, degrades model performance and leads to computational overhead. Due to these reasons, FS became an indispensable preprocessing step in fields dealing with high dimensional data. Traditional approaches evaluate features without considering the correlation among them and also this evaluation is performed on an individual basis. Furthermore, these methods generally fail to scale on a large space.

However, FS based on feature groping is a powerful approach since i) it discovers correlations among features by clustering ii) search dimension is significantly diminished iii) relieves computational burden. Although there are many papers dealing with this approach to a certain extent, to the best of our knowledge, none of them focus on this approach in detail as a review as stated here. For these reasons, we believe this paper will be more guiding and suggestive for those learning and working in deriving such methods compared to current literature.

Survey Methodology

Our main focus in this review is to examine FS approaches via clustering. In this context, we searched for the terms “feature grouping”, “feature selection based on grouping”, “attribute clustering” and “cluster-based feature selection” using Web of Science, Scopus, and Google Scholar. It is notable that we excluded those studies grouping samples (observations) or clustering features as the final outcome. We particularly point out that our fundamental focus is grouping of features as the preprocessing step followed by extraction of a reduced subset of features by a certain procedure which is subsequently input into a classification or clustering process. Studies of this paradigm under unsupervised setting are on a limited scale compared to supervised respect due to lack of labels in the former. Even though it’s not known clearly, we think that inclusion of this approach may have emerged late 90s. Recently, interest in this concept has grown rapidly with different forms as we shall see here. In fact, selection of significant features by removing irrelevant or redundant ones is just one aspect; ranking of these features in terms of being informative or having discriminative power, and stability of them for different models are other issues that are taken into consideration. Here, we examined different

studies during literature mining, categorized them, and presented readers a versatile work in which we aimed at providing a robust basis on the topic.

2. Basics of Feature Selection

In this section, we will present basic concepts in FS field. According to their interaction with classification model, FS techniques can be classified into filter, wrapper, and embedded techniques [7]. Later in the literature, hybrid and ensemble techniques have emerged as variants of them. Hybrid approach combines two different methods to utilize the advantages of both approaches where the common combination is filter and wrapper methods. Ensemble technique integrates an ensemble of feature subsets and then yields the result from the ensemble. The overview of the three main types of methods is shown in **Fig. I**.

2.1. Filter Method

Filter type methods select features by assessing intrinsic properties of data based on statistical measures instead of cross-validation performance. They are easily scalable to high-dimensional datasets, independent of the learning algorithm, simple and fast computationally, and resistant to overfitting. In this method, each feature is assigned a score determined by the statistical measurement selected. Afterwards, all features are ranked in descending order and those with low scoring are removed using a threshold value. The remaining features comprise the feature subset and are then fed into the classification model. Consequently, feature selection is carried out once and then various classifiers can be employed. Disadvantages of this technique are that features are selected irrespective of the classifier, and that feature dependencies are ignored. Some common statistical measures used in this technique are Information Gain (IG), Pearson's Correlation (PS), Chi Square (χ^2) and Mutual Information (MI).

2.1.1. Information Gain

Information gain (IG) is an entropy-based feature selection method and used to measure how much information a feature carries about the target variable. IG of a feature X, $IG(X)$, is calculated using

$$IG(X) = E(D) - \sum_{i=1}^{n} \frac{D_i}{D} E(D_i) \quad (1)$$

where $E(D)$ denotes the general entropy belonging to class labels, $\frac{D_i}{D}$ is the ratio of number of occurrences of each value on feature X, and $E(D_i)$ specifies the entropy of ith feature value calculated by splitting dataset D based on feature X.

2.1.2. Pearson's correlation

Pearson's correlation is a measure of the dependency (similarity) of two variables and used for finding the relationship between continuous features and the target feature. It has a value ranging between -1 to 1, where 1 shows a strong correlation and -1 means a total negative correlation. So, 0 value implies no correlation between the features. This method can also be used to measure correlation on a feature – feature basis in order to remove redundant features. Pearson's correlation coefficient can be found for feature X with values x and classes Y with values y where X, C are random variables by the following equation:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \quad (2)$$

2.1.3. Chi Square

Chi square (χ^2) is a statistical method to test the independence of two events. It's a measurement of the degree of association between two categorical values. It measures the deviation from the expected frequency assuming the feature event is independent of the class label. This assumption is tested by the formula

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (3)$$

where O_{ij} is the observed (actual) value and E_{ij} refers to the expected value suggested by the null hypothesis. Higher value of X^2 shows rejection to the null hypothesis, namely, higher dependency between the feature and the class label.

2.1.4. Mutual Information

Mutual information (MI) is another statistical method used to assess the mutual dependence between the two variables. MI quantifies the amount of information that one random variable includes in the other random variable. MI between two continuous random variables X and Y with their joint probability functions p(x,y), and their marginal probability density functions p(x) and p(y), respectively is given by

$$I(X;Y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad (4)$$

For discrete random variables, the double integral is substituted by a summation as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (5)$$

2.2. Wrapper Method

In this methodology, a search strategy for possible subsets of features is defined, and the learning algorithm is trained using these subsets in an iterative manner. Unlike filter methods, wrapper methods are in interaction with classifier, however, the evaluation of feature subsets is obtained using a specific classification model which makes this method specific to a learning model. Several possible combinations of features are evaluated in the model by wrapping the search algorithm around it. This method provides suboptimal feature subsets for training the model since evaluating all possible subsets is computationally not practical, and generally gives better predictive accuracy than filter methods but is computationally intensive due to searching overhead and learner dependence.

The search for generating subsets may be in a search space with schemes such as Forward Selection, Backward Elimination, Stepwise Selection or a heuristic search. Forward selection is a repetitive technique where no feature is considered onset. Initially, the feature with the best performance is added. Then another most significant feature giving the best performance together with the previously added feature is selected. This process proceeds until the inclusion of a new feature does not improve the classifier performance. In backward elimination, we begin with all the features available and discard the most insignificant feature from the model recursively. This elimination process is repeated until removal of features does not enhance the performance of the model. For stepwise selection, this technique is a combination of both forward selection and backward elimination. It starts with an empty set and the most significant feature is added at each iteration. While adding a new feature previously selected features are removed if any of them has become insignificant. Heuristic search is concerned with optimization and aims at optimizing the objective function in evaluation of different subsets.

Support Vector Machines with Recursive Feature Elimination (SVM-RFE) is a popular example of wrapper methods. The idea is mainly to train the classifier by the given data and assign a rank by SVM for each feature as its weight. Then, features with the smallest weights are removed by a specific rate determined by the user. This procedure is repeated until reaching a predefined number of features.

2.3. Embedded Method

This method includes advantages of filter and wrapper methods and performs feature selection and model construction at the same time. Just like wrapper techniques, they are specific to a learning model but they have less computational complexity than wrapper methods. One technique of this type of feature selection is regularization that adds a penalty to the coefficients

to overcome overfitting in the model. These coefficients may even be reduced to 0 such as in LASSO for features that do not contribute to the model. Features with non-zero coefficients are retained and those with low or zero coefficient are excluded. Another technique to integrate feature selection in model creation is decision trees. These tree-based methods are non-parametric models that consider features as nodes. Tree-based strategies used by random forests accumulate various numbers of decision trees and rank the nodes (features) by decrease in the impurity (Gini impurity) over all the trees.

3. Feature Selection Approaches

Broadly speaking, FS algorithms conducted in many studies can be categorized as two classes: i) traditional feature selection, ii) feature selection based on grouping. Whereas traditional approaches generally consider all features contingent on "singularity" during selection process, to put it another way, comprise inclusion or elimination of features based on some statistical measures or classifying capacity at a singular level, cluster-based methods, on the other hand, remove redundant and detect relevant features by grouping them into clusters, leading to reduced search space, too.

3.1. Traditional Feature Selection

Different FS methods exist in abundance in the literature, including filters based on distinct criteria (dependency, information, distance and consistency [8]), and wrapper and embedded methods employing different induction algorithms. Due to their simplicity, filter methods are often preferable in the context of high dimensional data; the absence of necessity for a search route and of interaction with classifier makes them computationally efficient and practically feasible in applications. A comparative work on various filtering methods (mixture model, regression modeling and t-test) was proposed in [9] and they outlined similar and dissimilar aspects of these methods. Lazar et al. [10] also reviewed filter type FS algorithms used in gene expression microarray analysis.

Wrapper methods carry computational burden due to requirement of navigation in search domain and to interaction with the predictor. However, they provide better accuracy than filters on account of interaction with learning algorithm. Talavera L. et al. [11] conducted their study making a comparison of filter and wrapper approaches in clustering. A recent study [12] suggests the overview of existing wrapper techniques including the pros and cons of them. Embedded methods, like wrapper techniques, possess computational complexity when it comes to high-dimensional data. They are more efficient than wrappers and have less complexity. Applications in bioinformatics under this approach has been reviewed in [13].

Hybrid methods combine two methods (such as filter and wrapper) to take advantage of both methods in order to increase efficiency and performance. Ensemble methods integrate different methods for FS, classification or both. In this approach, multiple feature selectors, induction algorithms, different subsets may be included according to the design scheme. A detailed

discussion on hybrid and a good review on ensemble feature selection techniques can be found in [14] and [15] , respectively. In some studies, FS methods are divided into these five categories [16].

Traditional FS approaches have several shortcomings. For instance, filter methods evaluate the significance of each feature individually without considering the relationships and interactions between the features. Wrapper methods can provide the optimal feature subset but their complexity makes them imperfect, they are not preferable especially in combinatorial techniques such as in ensemble methods. In addition, they are not applicable to data with small number of samples due to overfitting. Embedded methods, like wrappers, are specific to the model hence may give a different feature subset for the same dataset. The main drawback behind such methods is their inability to remove redundant measures and to retain informative features efficiently.

3.2. Feature Selection Through Feature Grouping

In this section, we will categorize FS approaches based on feature grouping under supervised and unsupervised context. Supervised FS utilizes data labels to measure importance and relevance of features. Unsupervised FS, on the other hand, assesses feature relevance by exploiting natural structure of the data without class label. As we shall see, a typical scenario in feature selection approaches based on grouping is that the features are first partitioned into clusters and then (a) representative feature(s) is (are) selected from each cluster according to a specific metric or technique as shown in **Fig. 2**.

3.2.1. Feature Selection under Supervised Setting

There are many studies conducted on FS through feature grouping in a variety of papers in the literature. The grouping of features is performed by various techniques including K-means, hierarchical clustering, graph theories, information theory metrics, kernel density estimation, logistic regression and regularization methods. With the availability of class labels in datasets, this prevalence is increasing day by day, offering new approaches and gaining new insights into the field.

Many diverse studies were carried out that performed K-means or hierarchical clustering for grouping features and then chose genes from each cluster. Sahu et al. [17] proposed an ensemble approach where K-means is applied first for feature grouping and then three different filter based ranking techniques (t-test, signal-to-noise ratio and SAM) are implemented for each cluster independently and the feature in the front rank from each cluster is selected to form three distinct feature subsets. Afterwards, subsets are applied additional elimination by checking the availability of each feature in a subset in another subsets. That is, a feature is discarded if it is not available in other subsets. Another study [18] applied information compression index to group features by hierarchical clustering and then sorted features within each cluster by Fisher criterion measuring the classifying capacity of each feature in a cluster. Subsequently, the feature in the

front rank is selected for each cluster to form the feature subset.

Au et al. [19] proposed an effective algorithm applied on gene expression data, called ACA, which uses an information measure to quantify correlation between features, and performs K-mode algorithm, similar to K-means, to cluster features. They defined mode of each cluster as the attribute (feature) with the highest sum of relevancy with others in each feature group. These modes constituted the final reduced subset. Their measure was also utilized to get good clustering configurations automatically. Chitsaz et al. [20] presented a fuzzy variant of this study which relies on the basic underlying idea in fuzzy clustering approaches, that each feature may belong to more than one group. Rather than considering association of each feature with a sole cluster, association with all features among the overall clusters is considered by assigning different grades of membership to features. Their extended work [21] integrates chi-square test to assess the dependency of each feature on the class labels during FS process.

Graph-based approaches are also common in studies involving FS through grouping. Song et al. [22] proposed an algorithm, called FAST, and benefited minimum spanning trees (MST) to create feature clusters. They adopted symmetric uncertainty to determine relevance between any pair of features or between the feature and the target class. Another study [23] under supervised framework similarly used MST for grouping and variation of information for relevance measure. Desired number of features and the pruning rate should be given as input in their algorithm. A quite recent study by Zheng et al. [24] builds the graph by interaction gain , makes use of MST to produce feature groups and probabilistic consistency measure for quality metric including two different techniques for FS: in the first one, they applied the conventional way of selecting representatives from each feature groups, and used harmony search as a metaheuristic search in the second. The metaheuristic approach dominates their first proposed algorithm together with other search mechanisms. Speaking of metaheuristic, Torres et al. [25] employed Markov blanket for clustering features and then these predominant groups are involved in Variable Neighborhood Search metaheuristic.

Although many studies focused their attention on discriminative power and redundancy removal of features, most of them neglect the stability of the selected features. Yu et al. addressed this issue in their two studies [26], [27]. In [26], rather than typical clustering algorithms, they applied kernel density estimation accompanied by iterative mean shift procedure to find feature clusters. Subsequently, these feature clusters were evaluated according to relevance using F-statistic and a representative feature from within each cluster is selected. The same authors extended this study in [27], where consensus feature groups were identified in an ensemble learning manner and features were extracted in the same way as their first study. They showed the stability of selected features by their algorithm in their experiments in both studies.

All the works mentioned until now are considered as global FS, i.e., finding a reduced subset of global features for the entire population. However, there are cases where these approaches are not applicable. For instance, take an image recognition task, where feature importance may alter since a set of relevant features may be important for identifying a specific object but insignificant for another object at a different position. This gap paved the way for a different technique, called Instance-wise FS that associates each feature’s relationship to its labels by assigning a different selector for each instance. Interested readers to grouping and selection of features in this

approach can refer to [28], [29]. A summary of approaches under supervised framework is outlined in **Table I**.

FS approaches based on clustering are not necessarily in the manner of grouping features into clusters and choosing representatives. Distinctly, selection of the features may happen with different cluster configurations. Moshlei et al. [33] initially implement K-means for clustering all samples for a given dataset and a sample from each cluster is chosen at random to acquire the samples with the greatest differences for the preliminary dataset. Subsequently, variances of all features on the determined samples are calculated and a predefined number of features with the highest variances are selected, thereby forming the primary dataset. Thereafter, remaining features are added gradually to this dataset and K- means clustering (with a predefined number of clusters) is applied iteratively in each step. Features causing changes in the structure of clusters are observed in a repetitive manner and considered as significant. Other features that don't lead to any alteration in clusters are eliminated.

Another work by Yousef et al. [34] gained *"recursive cluster elimination*" term into the community and their approach is adopted in many studies. Since this approach was widely employed by different studies, we elaborate this method in detail by reviewing its application areas and modified usages in Section 4.

3.2.2. Feature Selection under Unsupervised Setting

As with the traditional methods in FS, many of feature grouping-based FS approaches are in the supervised learning paradigm. Unsupervised FS is more challenging than supervised FS because of no prior knowledge about class labels and unknown number of clusters. Unsupervised FS methods typically involve 1) maximization of clustering performance by some index or 2) selection of features based on dependency. Since this paper is about FS, first one is out of scope of this study. Many statistical dependency/distance measures are available in the literature including correlation coefficient, least square regression error, Euclidean distance, entropy, and variance. Selected features in unsupervised FS methods can be evaluated in terms of both classification performance and clustering performance. **Table II** summarizes works on unsupervised FS based on clustering.

Mitra et al. [35] proposed an unsupervised feature selection algorithm using feature similarity. A new similarity measure called *maximum information compression index* is introduced in their study. Also, they demonstrated use of representation entropy for measuring redundancy and information loss quantitatively. Features are partitioned into clusters using k-NN principle along with a similarity measure. Entropy metric is chosen as the feature selection criterion and applied to select a single feature from each cluster to constitute the reduced subset. To evaluate the effectiveness of selected features, the proposed method is compared with KNN, Naive Bayes and class separability (including Relief-F) for classification capability , and with entropy and fuzzy feature evaluation index for clustering performance. Their algorithm is rapid since no search is required and also this study is one of the states of the art work in the literature.

Another example is the study of Li et al. [36], which uses the same similarity measure in [35] and employs a distance function to obtain clusters of features. A representative feature, having

the shortest distance to others within a cluster, is selected from each cluster. Their approach is based on hierarchical clustering which enables them to choose feature subsets with different sizes by choosing from top clusters in the hierarchy. Their algorithm works for both unsupervised and supervised learning. Moreover, they are doing clustering just one time in their algorithm. They presented their experimental results for both clustering and classification.

As stated previously, FS methods developed under unsupervised framework does not utilize class label. As an example, Covões T.F. et al. [37] presents a comparative study of their approach with the algorithm proposed by Mitra et al [35]. Again, maximal information compression index is utilized to find clusters of features. Hereafter, they employed the simplified silhouette (SS) criterion to find optimum clusters, allowing to find the number of clusters as well. The computation for simplified silhouette depends only on obtained partitions, not dependent on any clustering algorithm. Hence, this silhouette is, in addition to determining the number of clusters automatically, capable of evaluating partitions acquired by any clustering algorithms. They employed the k-medoids algorithm along with the silhouette method in order to achieve optimum clusters. Then the corresponding medoid for each cluster is selected as the representative feature. The prerequisite for number of clusters known a priori in this algorithm has been overcome by SS since one can implement this algorithm for different values of number of clusters, and then select the best clustering according to the maximum value obtained in SS.

Another study under unsupervised framework is suggested in [38], where maximal information coefficient and affinity propagation are exploited for selection of features. Features are chosen as the centroid of each cluster in the final step. Although they present competitive results in classification with typical classifiers, no comparison is made for clustering.

FS methods developed under supervised framework can be an inspiration to unsupervised studies. For instance, Zhou et al. [39] developed an attribute (feature) clustering algorithm along with an FS method in an unsupervised manner. Apart from this, a recent hybrid work which is a combination of grouping and binary ant system can be found in [40].

4. Feature Grouping with Recursive Cluster Elimination

In the original framework, the first step in SVM-RCE is to group genes into clusters using K-means by which correlated gene clusters are identified. As the second step, SVM is used to score (rank) these clusters and finally clusters with low scores are eliminated. Remaining genes (features) in clusters are combined and then clustering along with SVM is applied iteratively until a predefined number of clusters are left. In each iteration, surviving genes are used for classification to measure the accuracy at each level. Interests in this method have grown rapidly over time and many studies conducted their research integrating this approach.

Weis et al. [41] presented a SVM-RCE-like approach where they included assessment of clusters collaboratively rather than evaluating clusters individually.The study of Deshpande et al. [42] utilized SVM-RCE (although they call it RCE-SVM in their paper) with small modifications for brain state classification.

Another study presented by Luo [43], in order to reduce the computational complexity of SVM-RCE, infinite norm of weight coefficient vector from the SVM model is applied to score each cluster instead of scoring clusters by cross-validation. Their results show considerable reduction in computation time while exhibiting comparative performance as SVM-RCE.

In the study associated with military service members, in addition to statistical significance test, SVM-RCE is used to classify individuals between PTSD, PCS + PTSD, and controls [44]. The features are connectivity paths acquired from 125 brain regions. In their experimental works, using SVM-RCE, they conclude that higher classification rate by 4% is achieved through imaging-based grouping than conventional grouping. Furthermore, imaging measures dominate non-imaging measures by 9% for both conventional and imaging-based groupings.

Jin et al. [45] conducted a similar study and adopted a modified version of SVM-RCE in their study of brain connectivity where the diagnostic label of a novel subject is tested whether it belongs to subjects with PTSD or healthy group . The connectivity features are measured from mean resting-state time series taken from 190 regions across the entire brain. They employ SVM_RCE in their experimental work to suggest that dynamic functional and effective connectivity gives higher classification results compared to their static counterparts.

Interestingly, Zhao et al. [46] applied SVM-RCE tool to the detection of expression profiles identifying microRNAs related to venous metastasis in hepatocellular carcinoma.

Chaitra et al. [47] conducted a study to identify biomarkers of autism spectrum disorder (ASD) using imaging datasets. They utilized SVM-RCE to assess the classification performance for three distinct feature sets consisting of connectivity features alone, complex network (graph) measures alone, and a feature set including both. Their accuracy results are not competitive; however, the emphasis is on assessing different feature sets, especially on the combined feature set.

5. Grouping Features with Domain Knowledge

Aforementioned FS approaches typically use some statistics and computational tools to group and select the features without any domain knowledge. However, specifically in bioinformatics, integration of biological knowledge is essential for better improvement in the process of gene selection and machine learning [48]. The general idea in integrating biological knowledge for FS is to first apply a biological grouping function for grouping the genes, and then give each group a rank by scoring them using a machine learning algorithm. Finally, genes in the top groups form the reduced subset of features.

An integrative approach presented by Qi and Tang integrates Gene Ontology (GO) annotations into gene selection process, where they start by finding discriminative score for each gene (feature) applying Information Gain (IG) and eliminating those with a score of zero [49]. The next step is to annotate these genes with GO terms. After that, the score of each term is calculated as the mean of discriminative scores of associated genes involved in the respective

term. The GO term with the highest score is determined and the most discriminative associated gene is selected and extracted. The steps including calculation of scores for GO terms and selection of next most informative gene is repeated until the final subset completion. Their comparative work with sole IG shows the effectiveness of GO integration in the gene selection process.

Another integrative approach, Support Vector Machines with Recursive Network Elimination (SVM-RNE), was proposed in [50], which was an extension of SVM-RCE. Similarly, genes are grouped into clusters by GXNA [51] and clusters with low score are eliminated at each iteration. The algorithm terminates when some predefined constraints on the number of groups are met.

In SoFoCles [52], genes are initially ranked by typical filter methods such as information gain, Relief-F or χ^2 and then a reduced subset of genes is created by a given threshold. Next, for each gene in the reduced subset, semantically similar genes from GO are determined. Finally, top semantically similar genes are selected to enrich the reduced subset. Experimental works conducted using SoFoCles reveal enhancement in classification results by integrating biological knowledge into gene selection.

Mitra et al. [53] adopted CLARANS for gene (feature) clustering via gene ontology (GO) analysis. The final reduced feature subset is composed of genes which were medoids of biologically enriched clusters. In their experiments, incorporation of biological knowledge enhanced classifier performance and reduced computational complexity. The same authors subsequently made use of a fuzzy technique, FCLARANS, to obtain clusters and selected representative genes from clusters by fold change [54].

The study suggested by Fang et al.[55] includes combination of both KEGG and GO terms with IG. The initial dataset is applied IG as filtering and then GO and KEGG annotations are explored for the remaining genes. As the next step, association mining is applied to this annotation information and the interestingness of the frequent itemsets is determined by averaging the original discriminative scores (from IG) of the involved genes. The final gene set is attained via the selection of the highest ranked genes from the top n frequent itemsets. They assessed their method using GO, KEGG, and both against IG and study of [49]. Despite the lower rate of improvement in the overall accuracy, they are able to achieve it with a significant reduction in the number of genes.

Raghu et al. [56] utilize KEGG, DisGeNET and other genetic meta information in their integrated approach. Two metrics, gene importance and gene distance, are computed in their framework. Importance score for each gene is calculated using DisGeNET, which is a public platform containing gene collections associated with diseases. Distance between genes is computed based on their chromosomal locations and associations to the same diseases. Both scores are then employed to compose gene sets with maximum relevance and diversity. Compared to variance-based techniques, their method performs better in predictive modeling task on a small scale.

Perscheid et al. [57] makes a comparison between traditional and knowledge-based gene selection methods applied on gene expression data. Their approach produces gene rankings by

integrating knowledge bases and each of these rankings are evaluated with a predefined number of selected genes. Finally, the ranking with the best performance is selected. Moreover, they proposed a framework allowing external knowledge utilization, gene selection and evaluation in an automatic fashion. Although the framework seems to be knowledge base dependent, their experimental results demonstrate that incorporating biological knowledge into gene selection process upgrades performance in classification, decreases computational runtime, and enhances stability of selected genes.

Yet another study developed maTE [58], where gene groups are produced based on the miRNA target information and then each group is ordered by cross-validation. The average accuracy after a specific number of iterations determines the rank of each cluster. Genes on the top m groups are selected as the reduced subset.

The integrative FS method through grouping proposed by Yousef et al. [59] benefits from the biological knowledge for ranking and classification steps. Their proposed framework, named CogNet, initially implements pathfindR [60] to group the genes for clustering. These cluster groups are actually enriched KEGG pathways as a result of enrichment analysis. Then, a new dataset involving genes for the specific pathway is created for each cluster (pathway). These datasets are scored through Monte Carlo cross-validation (MCCV) and pathways are ranked according to the assigned scores. Ultimately, genes found in chosen top pathways are taken as features and used for classification.

Another study, called miRcorrNet [61], finds gene groups on the basis of their correlation to miRNA expression. Afterwards, these groups are applied a rank function for classification. The results showed AUC above 95% and that miRcorrNet is capable of prioritizing pan-cancer-regulating high-confidence miRNAs.

Very recently Zhang et al. [62] proposed a method DCG-Net; they quantify distance correlation gain between features to construct the biological network. In their algorithm, a greedy search method is applied to detect network modules. The edge with the highest weight is selected, then this edge is extended with respect to correlation metric to obtain the module in the network. This is done iteratively to extract modules and the module with the highest distance correlation is selected for analysis. Their experimental results showed effective results in terms of feature selection and classification accuracy.

6. Conclusions

The advances in high-throughput technologies has generated large high-dimensional data sets in many applications. The inevitable presence of redundant and noisy features increases computational complexity and degrades classifier capability. Hence, FS has become a required pre-processing step in itself as a primary concern for a long time. Here we present works done in the literature regarding FS techniques through feature grouping. Feature grouping is a powerful and efficient concept; it reduces search space and complexity, is resistant to the variations of samples, gives lower levels internal redundancy and provides better generalization capability of

the classifier. The form of feature grouping and selection of features out of groups are determined by different metrics or techniques as illustrated here.

During feature grouping, the aim is to keep similar features together within clusters while maximizing diversity between clusters. Different clustering algorithms exist and one needs to make sure for the quality of these clusters in the initial step. Choosing representative features or discarding less contributing clusters out of groups is another challenge. In fact, availability of independent and relevant features, correlation between features, and feature correlation to the decision are important items to be taken into consideration. More quality in terms of clusters and selection of genes, more informative and discriminative features in the reduced set.

In this study, our goal is to inform interested readers about trends in FS by feature clustering. Despite the wealth of many techniques in this field, there is still need for enhancement and novelty in the area. We believe approaches mentioned here may provide new insights into designing new schemes for FS in terms of better efficiency, effectiveness, stability, generalization and discrimination.

References

[1] B. Venkatesh and J. Anuradha, "A Review of Feature Selection and Its Methods," *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, Mar. 2019, doi: 10.2478/cait-2019-0001.

[2] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, May 2015, pp. 1200–1205. doi: 10.1109/MIPRO.2015.7160458.

[3] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," in *Machine Learning Proceedings 1994*, Elsevier, 1994, pp. 121–129. doi: 10.1016/B978-1-55860-335-6.50023-4.

[4] X. Shen and H.-C. Huang, "Grouping Pursuit Through a Regularization Solution Surface," *J. Am. Stat. Assoc.*, vol. 105, no. 490, pp. 727–739, Jun. 2010, doi: 10.1198/jasa.2010.tm09380.

[5] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, "Clustering approaches for high-dimensional databases: A review," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3, May 2019, doi: 10.1002/widm.1300.

[6] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[7] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997, doi: 10.1016/S0004-3702(97)00043-X.

[8] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 1–4, pp. 131–156, 1997, doi: 10.1016/S1088-467X(97)00008-5.

[9] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, no. 4, pp.

546–554, Apr. 2002, doi: 10.1093/bioinformatics/18.4.546.
[10] C. Lazar *et al.*, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 4, pp. 1106–1119, Jul. 2012, doi: 10.1109/TCBB.2012.33.
[11] L. Talavera, "An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering," in *Advances in Intelligent Data Analysis VI*, vol. 3646, A. F. Famili, J. N. Kok, J. M. Peña, A. Siebes, and A. Feelders, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 440–451. doi: 10.1007/11552253_40.
[12] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 International Conference on Engineering & MIS (ICEMIS)*, Agadir, Morocco, Sep. 2016, pp. 1–5. doi: 10.1109/ICEMIS.2016.7745366.
[13] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Brief. Bioinform.*, vol. 9, no. 5, pp. 392–403, Apr. 2008, doi: 10.1093/bib/bbn027.
[14] D. Asir, S. Appavu, and E. Jebamalar, "Literature Review on Feature Selection Methods for High-Dimensional Data," *Int. J. Comput. Appl.*, vol. 136, no. 1, pp. 9–17, Feb. 2016, doi: 10.5120/ijca2016908317.
[15] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Inf. Fusion*, vol. 52, pp. 1–12, Dec. 2019, doi: 10.1016/j.inffus.2018.11.008.
[16] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 13, no. 5, pp. 971–989, Sep. 2016, doi: 10.1109/TCBB.2015.2478454.
[17] B. Sahu, S. Dehuri, and A. K. Jagadev, "Feature selection model based on clustering and ranking in pipeline for microarray data," *Inform. Med. Unlocked*, vol. 9, pp. 107–122, 2017, doi: 10.1016/j.imu.2017.07.004.
[18] Z. Shang and M. Li, "Feature Selection Based on Grouped Sorting," in *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, Dec. 2016, pp. 451–454. doi: 10.1109/ISCID.2016.1111.
[19] Wai-Ho Au, K. C. C. Chan, A. K. C. Wong, and Yang Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 2, no. 2, pp. 83–101, Apr. 2005, doi: 10.1109/TCBB.2005.17.
[20] Chitsaz E., Taheri M., Katebi S.D., "A fuzzy approach to clustering and selecting features for classification of gene expression data". In: Proc. World Congress of Engineering (WCE 2008), 2008, 1650–1655.
[21] Chitsaz, E., Taheri, M., Katebi, S.D., Jahromi, M.Z.: An Improved Fuzzy Feature Clustering and Selection based on Chi-Squared-Test. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2009, Hong Kong, vol. I (2009)
[22] Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013, doi: 10.1109/TKDE.2011.181.
[23] Q. Liu, J. Zhang, J. Xiao, H. Zhu, and Q. Zhao, "A Supervised Feature Selection Algorithm through Minimum Spanning Tree Clustering," in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, Cyprus, Nov. 2014, pp. 264–271. doi: 10.1109/ICTAI.2014.47.

[24] L. Zheng, F. Chao, N. M. Parthaláin, D. Zhang, and Q. Shen, "Feature grouping and selection: A graph-based approach," *Inf. Sci.*, vol. 546, pp. 1256–1272, Feb. 2021, doi: 10.1016/j.ins.2020.09.022.
[25] M. García-Torres, F. Gómez-Vela, B. Melián-Batista, and J. M. Moreno-Vega, "High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach," Inf. Sci., vol. 326, pp. 102–118, Jan. 2016, doi: 10.1016/j.ins.2015.07.041.
[26] L. Yu, C. Ding, and S. Loscalzo, "Stable feature selection via dense feature groups," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, Las Vegas, Nevada, USA, 2008, p. 803. doi: 10.1145/1401890.1401986.
[27] S. Loscalzo, L. Yu, and C. Ding, "Consensus group stable feature selection," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, Paris, France, 2009, p. 567. doi: 10.1145/1557019.1557084.
[28] Q. Xiao, H. Li, J. Tian, and Z. Wang, "Group-Wise Feature Selection for Supervised Learning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 3149–3153. doi: 10.1109/ICASSP43922.2022.9746666.
[29] A. Masoomi, C. Wu, T. Zhao, Z. Wang, P. Castaldi, and J. Dy, "Instance-wise Feature Grouping," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 13374–13386. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/9b10a919ddeb07e103dc05ff523afe38-Paper.pdf
[30] S. Chormunge and S. Jena, "Correlation based feature selection with clustering for high dimensional data," *J. Electr. Syst. Inf. Technol.*, vol. 5, no. 3, pp. 542–549, Dec. 2018, doi: 10.1016/j.jesit.2017.06.004.
[31] C. H. Park, "A Feature Selection Method Using Hierarchical Clustering," in *Mining Intelligence and Knowledge Exploration*, vol. 8284, R. Prasath and T. Kathirvalavakumar, Eds. Cham: Springer International Publishing, 2013, pp. 1–6. doi: 10.1007/978-3-319-03844-5_1.
[32] R. Jensen, N. M. Parthalain, and C. Cornells, "Feature grouping-based fuzzy-rough feature selection," in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Beijing, China, Jul. 2014, pp. 1488–1495. doi: 10.1109/FUZZ-IEEE.2014.6891692.
[33] F. Moslehi and A. Haeri, "A novel feature selection approach based on clustering algorithm," *J. Stat. Comput. Simul.*, vol. 91, no. 3, pp. 581–604, Feb. 2021, doi: 10.1080/00949655.2020.1822358.
[34] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data," *BMC Bioinformatics*, vol. 8, no. 1, p. 144, Dec. 2007, doi: 10.1186/1471-2105-8-144.
[35] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002, doi: 10.1109/34.990133.
[36] Guangrong Li, Xiaohua Hu, Xiajiong Shen, Xin Chen, and Zhoujun Li, "A novel unsupervised feature selection method for bioinformatics data sets through feature clustering," in *2008 IEEE International Conference on Granular Computing*, Hangzhou, Aug. 2008, pp. 41–47. doi: 10.1109/GRC.2008.4664788.
[37] T. F. Covões, E. R. Hruschka, L. N. de Castro, and Á. M. Santos, "A Cluster-Based

Feature Selection Approach," in *Hybrid Artificial Intelligence Systems*, vol. 5572, E. Corchado, X. Wu, E. Oja, Á. Herrero, and B. Baruque, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 169–176. doi: 10.1007/978-3-642-02319-4_20.

[38] X. Zhao, W. Deng, and Y. Shi, "Feature Selection with Attributes Clustering by Maximal Information Coefficient," *Procedia Comput. Sci.*, vol. 17, pp. 70–79, 2013, doi: 10.1016/j.procs.2013.05.011.

[39] P.-Y. Zhou and K. C. C. Chan, "An unsupervised attribute clustering algorithm for unsupervised feature selection," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Campus des Cordeliers, Paris, France, Oct. 2015, pp. 1–7. doi: 10.1109/DSAA.2015.7344857.

[40] Z. Manbari, F. AkhlaghianTab, and C. Salavati, "Hybrid fast unsupervised feature selection for high-dimensional data," *Expert Syst. Appl.*, vol. 124, pp. 97–118, Jun. 2019, doi: 10.1016/j.eswa.2019.01.016.

[41] D. C. Weis, D. P. Visco, and J.-L. Faulon, "Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors," *J. Mol. Graph. Model.*, vol. 27, no. 4, pp. 466–475, Nov. 2008, doi: 10.1016/j.jmgm.2008.08.004.

[42] G. Deshpande *et al.*, "Recursive Cluster Elimination Based Support Vector Machine for Disease State Prediction Using Resting State Functional and Effective Brain Connectivity," *PLoS ONE*, vol. 5, no. 12, p. e14277, Dec. 2010, doi: 10.1371/journal.pone.0014277.

[43] Lin-Kai Luo, Deng-Feng Huang, Ling-Jun Ye, Qi-Feng Zhou, Gui-Fang Shao, and Hong Peng, "Improving the Computational Efficiency of Recursive Cluster Elimination for Gene Selection," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 1, pp. 122–129, Jan. 2011, doi: 10.1109/TCBB.2010.44.

[44] D. Rangaprakash *et al.*, "Compromised hippocampus-striatum pathway as a potential imaging biomarker of mild-traumatic brain injury and posttraumatic stress disorder," *Hum. Brain Mapp.*, vol. 38, no. 6, pp. 2843–2864, Jun. 2017, doi: 10.1002/hbm.23551.

[45] C. Jin *et al.*, "Dynamic brain connectivity is a better predictor of PTSD than static connectivity: Dynamic Brain Connectivity," *Hum. Brain Mapp.*, vol. 38, no. 9, pp. 4479–4496, Sep. 2017, doi: 10.1002/hbm.23676.

[46] X. Zhao, L. Wang, and G. Chen, "Joint Covariate Detection on Expression Profiles for Identifying MicroRNAs Related to Venous Metastasis in Hepatocellular Carcinoma," *Sci. Rep.*, vol. 7, no. 1, p. 5349, Dec. 2017, doi: 10.1038/s41598-017-05776-1.

[47] N. Chaitra, P. A. Vijaya, and G. Deshpande, "Diagnostic prediction of autism spectrum disorder using complex network measures in a machine learning framework," *Biomed. Signal Process. Control*, vol. 62, p. 102099, Sep. 2020, doi: 10.1016/j.bspc.2020.102099.

[48] M. Yousef, A. Kumar, and B. Bakir-Gungor, "Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data," *Entropy*, vol. 23, no. 1, p. 2, Dec. 2020, doi: 10.3390/e23010002.

[49] J. Qi and J. Tang, "Integrating gene ontology into discriminative powers of genes for feature selection in microarray data," in *Proceedings of the 2007 ACM symposium on Applied computing - SAC '07*, Seoul, Korea, 2007, p. 430. doi: 10.1145/1244002.1244101.

[50] M. Yousef, M. Ketany, L. Manevitz, L. C. Showe, and M. K. Showe, "Classification and biomarker identification using gene network modules and support vector machines," *BMC Bioinformatics*, vol. 10, no. 1, p. 337, Dec. 2009, doi: 10.1186/1471-2105-10-337.

[51] J. Wang *et al.*, "VISDA: an open-source caBIG™ analytical tool for data clustering and

beyond," *Bioinformatics*, vol. 23, no. 15, pp. 2024–2027, Aug. 2007, doi: 10.1093/bioinformatics/btm290.

[52] G. Papachristoudis, S. Diplaris, and P. A. Mitkas, "SoFoCles: Feature filtering for microarray classification based on Gene Ontology," *J. Biomed. Inform.*, vol. 43, no. 1, pp. 1–14, Feb. 2010, doi: 10.1016/j.jbi.2009.06.002.

[53] S. Mitra and S. Ghosh, "Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 1590–1599, Nov. 2012, doi: 10.1109/TSMCC.2012.2209416.

[54] S. Ghosh and S. Mitra, "Gene selection using biological knowledge and fuzzy clustering," in *2012 IEEE International Conference on Fuzzy Systems*, Brisbane, Australia, Jun. 2012, pp. 1–9. doi: 10.1109/FUZZ-IEEE.2012.6250797.

[55] O. H. Fang, N. Mustapha, and Md. N. Sulaiman, "An integrative gene selection with association analysis for microarray data classification," *Intell. Data Anal.*, vol. 18, no. 4, pp. 739–758, Jun. 2014, doi: 10.3233/IDA-140666.

[56] V. K. Raghu, X. Ge, P. K. Chrysanthis, and P. V. Benos, "Integrated Theory-and Data-Driven Feature Selection in Gene Expression Data Analysis," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, San Diego, CA, USA, Apr. 2017, pp. 1525–1532. doi: 10.1109/ICDE.2017.223.

[57] C. Perscheid, B. Grasnick, and M. Uflacker, "Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches," *J. Integr. Bioinforma.*, vol. 16, no. 1, Dec. 2018, doi: 10.1515/jib-2018-0064.

[58] M. Yousef, L. Abdallah, and J. Allmer, "maTE: discovering expressed interactions between microRNAs and their targets," *Bioinformatics*, vol. 35, no. 20, pp. 4020–4028, Oct. 2019, doi: 10.1093/bioinformatics/btz204.

[59] M. Yousef, E. Ülgen, and O. Uğur Sezerman, "CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis," *PeerJ Comput. Sci.*, vol. 7, p. e336, Feb. 2021, doi: 10.7717/peerj-cs.336.

[60] E. Ulgen, O. Ozisik, and O. U. Sezerman, "pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks," *Front. Genet.*, vol. 10, p. 858, Sep. 2019, doi: 10.3389/fgene.2019.00858.

[61] M. Yousef, G. Goy, R. Mitra, C. M. Eischen, A. Jabeer, and B. Bakir-Gungor, "miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking," *PeerJ*, vol. 9, p. e11458, May 2021, doi: 10.7717/peerj.11458.

[62] Y. Zhang, X. Lin, Z. Gao, and S. Bai, "A novel method for feature selection based on molecular interactive effect network," *J. Pharm. Biomed. Anal.*, vol. 218, p. 114873, Sep. 2022, doi: 10.1016/j.jpba.2022.114873.

Table 1(on next page)

Applications of FS by Grouping under Supervised Context

1 **Table I. Applications of FS by Grouping under Supervised Context**

Grouping Method	FS Method (metric)	Validation	Application Area	Study
K-means	correlation	classification accuracy	text and microarray	[30]
	ensemble	LOOCV	microarray	[17]
Hierarchical	Fisher	classification accuracy	miscellaneous	[18]
	Fisher	cross validation	miscellaneous	[31]
ACA	Interdependence mesure	classification accuracy	synthetic & gene expression	[19]
Fuzzy	correlation	classification accuracy	miscellaneous	[32]
	fuzzy	classification accuracy	miscellaneous	[21]
	fuzzy	classification accuracy	microarray	[20]
	probabilistic consistency	cross validation	miscellaneous	[24]

Graph-based	variation of information	silhoutte index & classification accuracy	miscellaneous	[23]
	symmetric uncertainty	classification accuracy	miscellaneous	[22]

2

Table 2(on next page)

Applications of FS by Grouping under Unsupervised Context

1 **Table II. Applications of FS by Grouping under Unsupervised Context**

Grouping Method	FS Method (metric)	Validation	Application Area	Study
Louvain community detection	binary ant system	classification error	Real-world datasets	[40]
k-mode	mode of each cluster	classification accuracy	miscellaneous datasets	[39]
Affinity Propagation	centroid of each cluster	classification accuracy	miscellaneous datasets	[38]
k-medoids	medoid of each cluster	classification accuracy	miscellaneous datasets	[37]
hierarchical	Feature with the shortest distance in the cluster	Minkowski Score	Gene datasets	[36]
kNN	entropy	entropy, fuzzy feature evaluation index, classification accuracy	Real life public domain	[35]

2

Figure 1

Typical approach for representative feature selection based on grouping

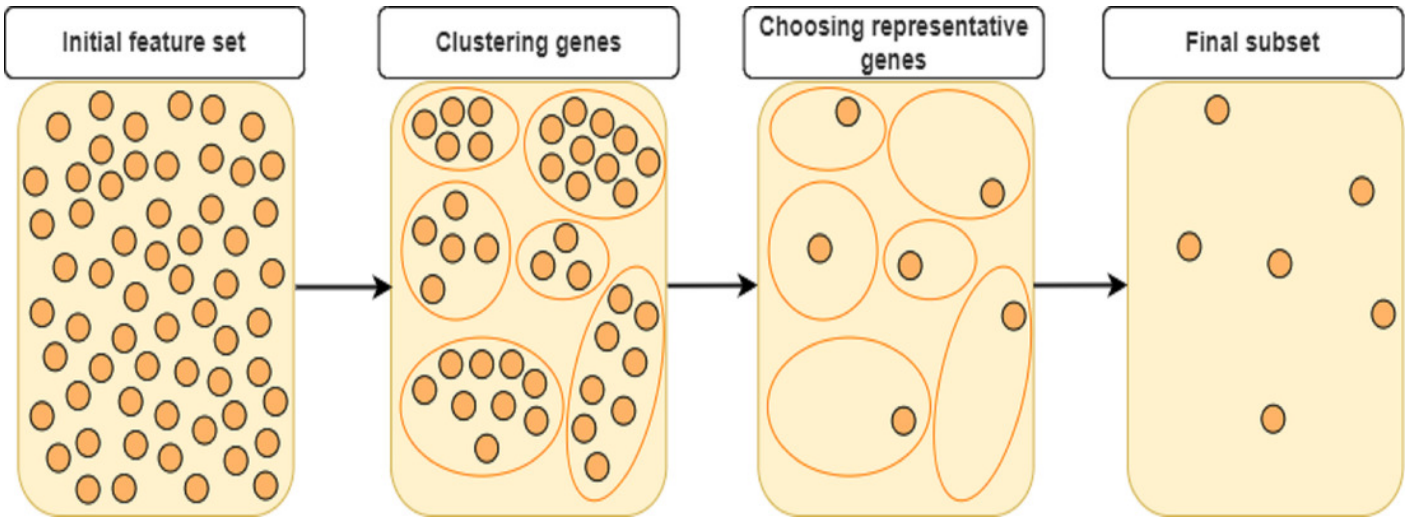

Figure 2

Three basic types of FS methods

'(A) Filter (B) Wrapper (C) Embedded.'