Artificial Intelligence Review
https://doi.org/10.1007/s10462-019-09800-w

A survey on feature selection approaches for clustering

Emrah Hancer[1] · **Bing Xue**[2] · **Mengjie Zhang**[2]

Abstract
The massive growth of data in recent years has led challenges in data mining and machine learning tasks. One of the major challenges is the selection of relevant features from the original set of available features that maximally improves the learning performance over that of the original feature set. This issue attracts researchers' attention resulting in a variety of successful feature selection approaches in the literature. Although there exist several surveys on unsupervised learning (e.g., clustering), lots of works concerning unsupervised feature selection are missing in these surveys (e.g., evolutionary computation based feature selection for clustering) for identifying the strengths and weakness of those approaches. In this paper, we introduce a comprehensive survey on feature selection approaches for clustering by reflecting the advantages/disadvantages of current approaches from different perspectives and identifying promising trends for future research.

Keywords Clustering · Feature selection · Data mining · Evolutionary computation

1 Introduction

With the developments of the computer hardware, software and online database technologies, it is now possible to collect and store large-scale datasets from different sources more than ever. The growth in the data concerning the dimensionality and the number of instances can also be illustrated from the UCI machine learning repository (Dheeru and Karra Taniskidou 2017). However, such growth has also led challenges in pattern recognition and knowledge discovery processes since not only the required information but also various noise may present in the data, due to the deficiencies of technological devices used

✉ Emrah Hancer
emrahhanc@gmail.com; ehancer@mehmetakif.edu.tr

Bing Xue
Bing.Xue@ecs.vuw.ac.nz

Mengjie Zhang
Mengjie.Zhang@ecs.vuw.ac.nz

[1] Department of Computer Technology and Information Systems, Mehmet Akif Ersoy University, Burdur, Turkey

[2] Evolutionary Computation Research Group, School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand

Published online: 02 January 2020

for the collection process or the nature of the dataset itself. For instance, in a scanned digital map, the background is eroded by the scanner when there exist two closely spaced objects. This kind of noise is the result of the nature of the data. Meanwhile, the point spread function of the scanner mostly affects the quality of the scanned map (Samet and Hancer 2012; Hancer et al. 2014). For another instance, noise, atmospheric conditions, constant offset, and slant offset are only some conditions that need to be handled before the further process of hyperspectral image processing (Awad 2018; Amini et al. 2018). In order to extract useful knowledge from such datasets, data preprocessing techniques are often required.

One of the well-known techniques in data preprocessing, *dimensionality reduction* tries to remove irrelevant and redundant features which degrade the performance of the further process (e.g., classification, regression, random forest, and clustering). Dimensionality reduction techniques can be categorized into feature extraction/construction and feature selection. In feature extraction/construction, the dimensionality of the data is reduced by deriving new features from the available original features. The representative approaches of feature extraction are Principal Component Analysis (PCA) (Jolliffe 1986), Linear Discriminant Analysis (LDA) (Ye 2007), and Singular Value Decomposition (SVD) (Golub and Reinsch 1970), just to name a few. Feature selection, on the other hand, tries to select a small subset of relevant (original) features from all available features according to a predefined evaluation criterion. In general, feature selection is needed when there is a large number of features in the dataset, and when users would like to build generalizable models, decrease the computational complexity, and reduce the required storage. These can also be achieved by feature extraction, but feature selection is more needed when the readability and interpretability of the models or data are important since the originality of features are kept in the reduced space (Hancer et al. 2018).

In the term of supervised learning, a feature is referred to as relevant or highly discriminant if it can discriminate instances that belong to different classes. In other words, feature selection approaches often require the class labels to determine whether a feature is relevant or not. However, not only in supervised tasks but also in unsupervised tasks high dimensionality may adversely affect the performance of the learning process and generally remains intensive computational costs. How can a feature be defined as relevant when the class labels are unknown, i.e., in the terms of unsupervised learning? It is not possible to make a clear definition of relevancy in the terms of unsupervised learning. However, it is suggested in (Li et al. 2016) that feature selection may also enhance the performance of an unsupervised learning algorithm from different perspectives as in supervised learning, such as reduce the learning time, simplify the learned prototype/model, and help the interpretability and visualization of the data.

One of the most typical tasks in unsupervised learning, clustering is the process of dividing instances into groups or clusters using a certain similarity criterion. Each cluster is expected to have maximal homogeneity within the cluster and maximal heterogeneity between clusters. In recent years, a variety of feature selection approaches have been designed and utilized for clustering, although as common as for classification. In the literature, some surveys have been provided to identify the issues involved in these approaches. Alelyani et al. (2013) considered feature weighting approaches for clustering. In this review, the authors preferred to thoroughly explain the overall schedule of approaches rather than reflecting the weaknesses and strengths of approaches. Amorim (2016) specifically focused on feature selection approaches wrapping the K-means algorithm and introduced a comparative study of these approaches. Mugunthadevi et al. (2011) reviewed feature selection approaches proposed for document clustering. However, this review did

not provide comprehensive analysis and discussions concerning this issue. In another one, Liu and Yu (2005) reviewed feature selection algorithms for both classification and clustering and proposed a taxonomy of feature selection approaches in a three-dimensional framework. Unfortunately, this taxonomy could not be treated as general since only a small number of studies were considered to build the taxonomy, especially for clustering. Miruthula and Roopa (2015) put forward a study on feature selection approaches such as linked unsupervised feature selection, spatial-spectral feature selection and sparse regression. Li et al. (2016) reviewed feature selection approaches proposed for clustering from data perspective, such as streaming data, link data, and text data. Another recently introduced study (Solorio-Fernandez et al. 2019) reviewed a variety of unsupervised feature selection approaches used in both clustering and classification tasks, and then conducted a comprehensive comparative analysis of several feature selection algorithms. It could be treated as successful in terms of providing new insights from different perspectives compared to the above mentioned surveys, but a variety of wrapper approaches were missing. In summary, most of the existing and recently proposed studies carried out in this field are missing in these surveys, and it is not possible to make a comprehensive analysis from all different perspectives. In particular, feature selection approaches designed for evolutionary clustering have not been fully considered and discussed in the literature yet, to the best of our knowledge. To alleviate this challenge, we introduce a comprehensive survey of feature selection approaches for clustering with a brief discussion on open issues and future trends. The target readers of this survey also include researchers who are interested in feature selection in general, and researchers, such as Ph.D. students who would like to have an overview of various existing methods in this area. Furthermore, this survey also aims to attract the interest of researchers, who usually focus on developing feature selection approaches for classification, to pay much attention to design feature selection approaches to address clustering problems.

The rest of this paper is organized as follows. Section 2 provides background on feature selection and clustering. Section 3 reviews feature selection approaches designed for clustering. Section 4 expresses the open issues of approaches with future trends. Section 5 concludes the paper with general conclusions.

2 Background

2.1 Feature selection

Feature selection is the process of removing irrelevant and redundant features, which have a detrimental effect on model construction. A predefined criterion evaluates the optimality of the selected feature subset. As feature selection keeps the original meaning of the original features and so provides better readability and interpretability, it has been widely used in a variety of data mining and machine learning tasks such as regression, classification and association rules. In the past, researchers mainly tried to explain and analyze feature selection on supervised learning tasks, especially classification due to the popularity of feature selection in such fields (Dash and Liu 1999; Xue 2014). Feature selection can be fundamentally considered in two scenarios: supervised feature selection and unsupervised feature selection. When feature selection is applied to unsupervised tasks (e.g., clustering), the general procedure of feature selection needs to be reconsidered. Typically, a feature selection approach involves four stages, namely

selection, evaluation, stopping criterion and validation (Liu and Yu 2005). In the first stage, a feature subset is selected using a predefined search strategy, such as complete search, sequential search and sequential floating search. In the second stage, the selected feature subset will be evaluated based on a certain criterion. After the stopping criterion is met, the subset which has the best evaluation value from all the possible subsets is chosen. Finally, the chosen subset is validated using validation metrics. In both supervised and unsupervised concepts, feature selection approaches can be broadly categorized as follows:

- *Filters* try to select an optimal feature subset according to the general characteristics of the data instead of a learning algorithm. Typically, filters calculate the score of a feature (subset) according to certain evaluation criteria. Then, they choose the features with the best scores. The evaluation criteria may be multivariable or univariable measures. While multivariable measures consider more than two-way relationships within the feature set, univariable measures evaluate each feature independently. Accordingly, multivariable measures can detect redundant features and thereby is treated as more general.
- *Wrappers* require a learner to evaluate the goodness of possible feature subsets. Wrappers can, therefore, achieve better feature subsets to enhance the performance of the predefined learning algorithm, but they tend to be computationally more intensive than filters. First, wrappers obtain a feature subset using search strategies. Second, the quality of the selected feature subset is evaluated through a learning algorithm. This procedure is repeated until the stopping criterion is met.
- *Hybrid* approaches aim to get the advantages of both wrappers and filters. Two hybridization ways are often used to hybridize wrappers and filters together. One way is to apply a two-stage process in which a filter approach is performed to reduce the feature set and then a wrapper approach is carried out on the reduced set to obtain the final subset. The other way is to use a filter (wrapper) approach as a local search mechanism in a wrapper (filter) approach. The latter way is expected to achieve better performance in terms of the learning performance and the feature subset size.
- *Embedded* approaches, referred to as a trade-off between wrappers and filters, embed feature selection into the process of a learning algorithm. Thus, they utilize the characteristics of wrapper and filter methods. First, they are in cooperation with the learning algorithm as in the wrappers. Second, they do not need to use the learning algorithm several times and thereby are far more efficient than the wrappers. However, embedded approaches often cannot reach better learning performance than wrappers.

In general, most filter feature selection methods are good for categorical datasets since filter measures, such as mutual information, work mainly on discrete variables. If the dataset contains continuous variables or mixed types of variables, wrapper methods might be better, since there are many different types of learning algorithms that can choose from to handle the data. The use of embedded method is highly depends on the specific learning algorithm's ability in handling categorical and/or continuous data. The hybrid methods are often used when the dataset has a very large number of features and many irrelevant features. Furthermore, the dataset with a small number of records is generally much harder, i.e. not only needs to consider which feature selection method to, but also need to consider the design process of how to apply the feature selection method to avoid overfitting or feature selection bias. However, the best method for a given dataset is of course the one with fine tuned parameters and specifically designed with a specific preprocessing step.

2.2 Clustering

Clustering is the process of automatically grouping unlabelled instances in a given dataset using predefined similarity measures, such as Euclidean distance, point symmetry and hamming coding. Similar instances are determined as a cluster. In particular, each cluster is represented via its cluster centroid. Clustering has been widely applied to a variety of data mining and machine learning tasks including image analysis, pattern recognition, information retrieval and wireless networks, just to name a few.

A great number of clustering approaches have been proposed in the literature. These approaches can be broadly categorized into partitional, hierarchical, density-based and model-based approaches (Hancer and Karaboga 2017). Partitional approaches iteratively allocate instances to clusters using distance-based metrics. These approaches tend to produce one level and non-overlapping spherical shapes. The most representative approaches belonging to this category are K-means (Macqueen 1967), K-medoids and Fuzzy C-means (FCM) (Bezdek et al. 1984). Hierarchical approaches try to build a hierarchy by partitioning instances into different levels. The hierarchy can be built through merging or splitting clustering. The representative methods of this category are BIRCH (Zhang et al. 1997), CURE (Guha et al. 1998), DIANA (Patnaik et al. 2016) and ROCK (Guha et al. 1998). Density-based approaches try to detect clusters by considering the density of regions in the data. Low-density regions distinguish instances from different clusters. The most popular approaches of this category are DBSCAN (Ester et al. 1996), OPTICS (Ankerst et al. 1999) and DENCLUE (He et al. b). Model-based clustering tries to postulate a statistical model for the data and then use a probability derived from this model as the clustering criterion. The representative methods of model-based clustering are expectation-maximization (McLachlan and Krishnan 2008) and Gaussian mixture model (McLachlan and Krishnan 2008). For more information concerning clustering, please see (Hancer and Karaboga 2017; Jain et al. 1999; Rui and Wunsch 2005).

2.3 Fundamental clustering approaches

In this section, we will provide an overview of clustering approaches which will be considered in the next section.

K-means: Arguably the most popular partitional algorithm, K-means (Macqueen 1967) tries to iteratively allocate data instances into clusters by evaluating the Euclidean distance of each instance to another one. For a given dataset Z, the output of K-means is a disjoint set of clusters $C = C_1, C_2, \ldots, C_k, \ldots, C_K$, where each cluster C_k is represented by a cluster centroid $m_k \in M$, $M = \{m_1, m_2, \ldots, m_K\}$. The centroid m_k is expected to have the smallest sum of Euclidean distances within C_k. In other words, K-means iteratively tries to minimize the within-cluster distance between data instances ($z_p \in C_k$) and respective centroids ($m_k \in M$), defined as follows.

$$W(Z, C) = \sum_{k=1}^{K} \sum_{z_p \in C_k} \sum_{f_i \in F} (z_{pi} - m_{ki})^2 \tag{1}$$

where m_{ki} indicates the *ith* feature of the *kth* centroid (m_k), and F is the feature set representing the N number of features such that $F = \{f_1, f_2, \ldots, f_i, \ldots, f_N\}$.

The general schedule of K-means is as follows:

1. Initialize a set of clusters C by selecting instances from the dataset or randomly generating within the predefined boundaries,
2. Assign each instance z_p to its closest cluster m_k through the Euclidean distance,
3. Update each cluster centroid $m_k \in M$ by taking the mean of all instances within C_k,
4. Go to Step 2 and repeat the process until the specific requirements are met.

Model-based clustering: Mixture models extend cluster analysis available to the data analyst. Mixture models define the structure of a cluster within a probabilistic scheme. If clustering is designed using finite mixtures of multivariate Gaussian distributions, a probabilistic distribution is prespecified as a data-generating process for the observed data. In particular, the data in each cluster and the combined data stems are assumed to be respectively generated from a multivariate Gaussian distribution and a convex combination of multivariate Gaussian distributions (Grün 2019). The general distribution used in mixture models is given by:

$$f(z_i, \Theta) = \sum_{g=1}^{G} \pi_g f(z_i|\Theta_g) \tag{2}$$

where $z_i = (z_{i1}, z_{i2}, \ldots, z_{iM})$ is the M-dimensional vector representing the *ith* instance of Z, π_g denotes the mixing probability of the *gth* group and Θ_g denotes the parameter set of the *gth* group.

Compared with K-means, Gaussian mixture modelling allows clusters with different sizes and different volumes. Moreover, the clusters are independent of the scaling used for the variables (except for potential numerical issues). In summary, the general working concept of mixtures models is flexible but a bit complex.

Evolutionary clustering: As an innovative discipline of artificial intelligence, evolutionary computation (EC) techniques have been strongly believed to be effective metaheuristics for NP-hard problems. In particular, EC techniques (Zhao et al. 2017; Parvin et al. 2012) can provide optimal or near-optimal solutions in a reasonable time to such problems. EC techniques include mainly two categories: evolutionary algorithms and swarm intelligence algorithms. The former ones follow the fundamental biological terms such as mutation, recombination and selection. The representative methods for this category are genetic algorithms (GAs) (Holland 1975), genetic programming (GP) (Koza 1992), and differential evolution (DE) (Storn and Price 1997). The latter ones are inspired by the intelligent behaviours of insects or animals such as birds, ants, bees and bacterial swarms. The most popular examples for this category are particle swarm optimization (PSO) (Kennedy and Eberhart 1995), ant colony optimization (ACO) (Dorigo and Di Caro 1999) and artificial bee colony (ABC) (Karaboga et al. 2014).

As clustering was deemed as an NP-hard problem (Aloise et al. 2009), EC techniques have been widely used to develop clustering approaches, especially partitional approaches. For more information concerning EC techniques for clustering, please see (Hancer and Karaboga 2017; Hruschka et al. 2009). The generalized mechanism of an EC-based clustering approach can be described as follows:

1. Initialize each solution in the population with K cluster centroids.
2. Perform predefined evolutionary operators to each current solution to generate new solutions.

3. Assign data instances to each cluster centroid in the new solution using a predefined distance metric (e.g., Euclidean distance).
4. Evaluate the fitness value of the new solution by a predefined clustering criterion [e.g., Eq. (1)].
5. If the fitness value of a new solution is better than its current solution, the new solution is kept in the population instead of its current solution.
6. Repeat steps 2–5 until the maximum number of iterations or a predefined stopping criterion is met.
7. Choose the solution, which has the best fitness value in the population, as the final cluster centroids.

2.4 Principles of feature selection for clustering

From the perspective of clustering, irrelevant and redundant features in the data may degrade the quality of clusters, lead to high computational cost, and require more memory. Therefore, eliminating such irrelevant and redundant features may alleviate these issues. In order to illustrate this notion, we provide Fig. 1 (Li et al. 2016). According to Fig. 1, f_1 is sufficient to distinguish clusters. On the other hand, f_2 referred to being irrelevant does not have an additional effect on the clustering process, and f_3 referred to as redundant adversely affects the homogeneity of clusters. Moreover, different feature subsets comprising of relevant features may produce different clusters. For instance, a relevant feature may help to discover hidden patterns of the data when added to a feature subset. Accordingly, a variety of feature selection approaches have been proposed to utilize in clustering approaches in order to increase the clustering performance by eliminating irrelevant and redundant features.

Similar to feature selection in classification, feature selection approaches proposed for clustering can be split into filter, wrapper, embedded and hybrid approaches. However, there has been much less work on embedded and hybrid feature selection approaches in clustering. While wrappers are dependent on a clustering algorithm to evaluate the clustering quality of a selected feature subset, filters are independent of a clustering algorithm. Like wrappers, embedded approaches also work with a clustering algorithm but different from wrappers by incorporating knowledge related to the specific structure of clustering. Another type of methods are hybrid approaches that combine filter and wrapper approaches into a single strategy.

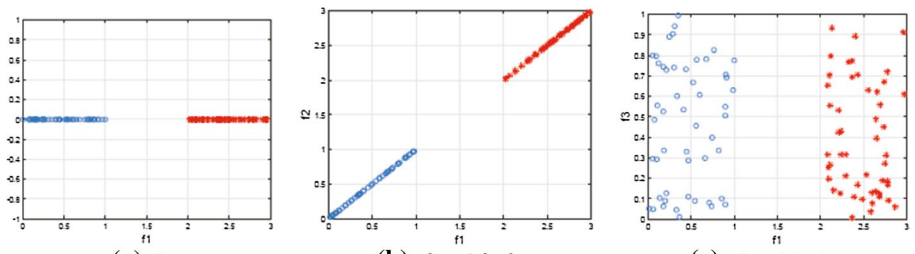

Fig. 1 An illustrative example to show how irrelevant and redundant features affect clustering (Li et al. 2016)

Treated as an extension of feature selection, subspace clustering attempts to find the cluster structure in different subspaces of the same dataset. Like feature selection, subspace clustering also requires a search strategy and an evaluation metric. In addition, subspace clustering must consider to limit the scope of the evaluation metric to consider different subspaces for each different cluster. In this paper, we do not deeply consider subspace clustering approaches to prevent adding a large number of pages in the study, but we will discuss some recent subspace clustering algorithms which are mainly in the category of wrapper approaches. For more information concerning subspace clustering, please see Parsons et al. (2004), Domeniconi et al. (2004).

3 Feature selection approaches for clustering

We investigate feature selection approaches designed for clustering in the four general categories, which are filter, wrapper, embedded and hybrid approaches. Since there are a lot more methods belong to the wrapper approaches, we will further divide wrapper approaches into subcategories which are feature selection approaches with K-means, feature selection approaches with model-based clustering and feature selection approaches with evolutionary clustering. Figure 2 shows the taxonomy of the approaches reviewed in this paper.

3.1 Filter approaches

Filters select features in the data according to the characteristics of features. In contrast to wrappers, filters do not require any clustering/learning algorithm and thereby are more efficient than wrappers. Instead of evaluating the clustering performance of the data through a clustering algorithm, filters directly evaluate the statistical performance of features in the data, which is more suitable for large-scale datasets. Therefore, the clustering performance of filters is generally lower than that of the wrappers. The number of works focusing on filters is much less than on wrappers in the literature.

A variety of criteria have been used to define the relevancy of features in the data. Among these criteria, some of them use data similarity to measure feature importance, referred to as similarity-based approaches (Li et al. 2016). In the term of supervised feature

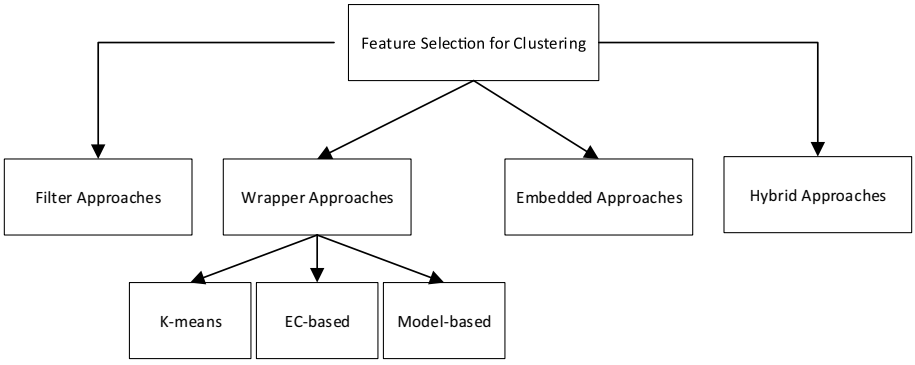

Fig. 2 Categorization of feature selection approaches for clustering

selection, data similarity can be obtained through the label information. On the other hand, in the term of unsupervised feature selection, all features in the data are individually evaluated through pairwise similarity among instances and then a feature subset, which well preserves the utility maximization, is selected.

Mitra et al. (2002) proposed a filter approach by measuring the dependencies between features based on a type of variance-based metric, referred to as the maximal information compression index (MICI). The overall aim of the proposed approach is to divide features into clusters using the principles of the k-nearest neighbor algorithm, i.e. features within the same cluster are similar to each other, while features in different clusters are dissimilar. In each iteration, the k nearest features are found for each feature based on the MICI criterion. Then, the feature, which builds the most compact subset with k nearest features, is selected. This procedure is repeated until all features are selected or discarded. The improved version of Mitra's work (Li et al. 2007) applies the feature selection approach in (Mitra et al. 2002) to remove redundant features and then uses an exponential entropy criterion to sort the remaining features in terms of their relevancy. Finally, a feature subset is selected using a forward search mechanism and the fuzzy evaluation index (FFEI).

He et al. (2005) introduced a filter approach (referred to as Laplacian Score) which aims to select features according to locality preserving power. Laplacian Score can be applied to both unsupervised and supervised problems. To test the performance of Laplacian Score, it was first applied to face image clustering. According to the results, it performed better than Variance Score which typically eliminates all zero variance features. However, the feature subset size should be prespecified by a user. Accepted as a particular case of Laplacian Score, the spectral feature selection (SPEC) algorithm (Zhao and Liu 2007) uses the radial-basis function as a similarity function to evaluate the feature relevancy. The relevant features are expected to have similar similarity values. The experimental results showed that it performed better than Laplacian Score.

Haindl et al. (2006) introduced a filter approach based on the Pearson correlation coefficient. This approach first calculates all possible pairwise correlations between features in the data and then removes the feature which has the highest average dependency with other features. This procedure is repeated for a number of features that is specified by a user. Another filter feature selection approach (called Relevance Redundancy Feature Selection (RRFS)) (Ferreira and Figueiredo 2012) designed for both unsupervised and supervised tasks involves two stages. For unsupervised scheme, in the first stage, features are ranked in terms of the variance criterion. In the second stage, the ranked features are evaluated in terms of redundancy using the variance criterion. Then, m features with the lowest redundancy are chosen for the feature subset.

Nie et al. (2008) introduced a filter approach that selects features using a scatter trace criterion. This approach constructs two similarity matrices to measure the within-cluster distance and the between-cluster separation. However, the scatter trace criterion does not include a closed form solution and thereby is difficult to be considered. To deal with this issue, the scatter trace criterion is often reformulated as a more suitable format, referred to as the ratio trace, e.g. Wang et al. (2007) developed a strategy to optimize scatter trace criterion iteratively.

Based on the scatter trace criterion, Li et al. (2008) proposed a localized feature selection approach. It is revealed in (Li et al. 2008) that the within-cluster distance and the between-cluster separation need to be normalized due to the following reasons. First, scatter trace tends to increase monotonically with dimensionality while no change is observed in clustering assignments. Second and last, this approach tries to choose different feature subsets for each cluster. To perform normalization on both the within-cluster distance and

the between-cluster separation, cross-projection is performed over an individual cluster set. In this approach, a cluster set is first generated over all available features and then a sequential backward selection is carried out to eliminate irrelevant and redundant ones. It can therefore be suggested that the computational cost may be extremely intensive for datasets with a large number of features.

Vandenbroucke et al. (2000) introduced a filter approach for the segmentation of soccer images. First, many texture color features are obtained from source images by considering pixel values through different color spaces such as RGB and CIE. Then, the original feature set is categorized into several feature subsets using a type of the scatter trace criterion. For each possible feature subset, the value of the scatter criterion is calculated, and the feature subset which maximizes the scatter criterion is selected to determine the candidate feature. If the candidate feature is highly correlated with the selected feature, it is abandoned. Although this approach produced great segmentation performance, it is particularly designed to segment soccer images and it was not tested on different kinds of image datasets.

3.2 Wrapper approaches

In contrast to a filter approach, a wrapper approach is designed for a specific clustering algorithm. In other words, wrappers synchronously work with clustering algorithms and thereby are often able to reach great clustering performance compared to filters. In this section, we present and discuss wrapper approaches specifically designed for K-means, model-based and evolutionary computation based clustering algorithms.

3.2.1 Feature selection for K-means

Despite its popularity, K-means also has deficiencies like as any other algorithm in data mining. First, the clustering performance of K-means is very dependent on the initial cluster set. Second, the number of clusters K needs to be predefined by a user. Third, it may converge to local minima since it performs greedy search. Fourth and last, it assumes all features are equally important rather than considering the actual degree of their relevance. Due to the overall goal of our study, we particularly focus on methods handling the last deficiency.

K-means evaluates each feature with equal importance. In other words, each feature even defined as less relevant, irrelevant or redundant has the same contribution to the clustering process. However, it is pointed in (Chakraborty and Das 2018) that if there exist a significant number of irrelevant and redundant features for the clustering process, the quality of clusters may be dramatically low. Moreover, K-means is very likely to be negatively affected from the presence of irrelevant and redundant features. This drawback has been addressed by assigning weights to each feature, referred to as feature weighting. Feature selection assumes all relevant features in the selected subset have the same degree, i.e., there exist only two cases such that a feature may be selected or not for the feature subset. Feature weighting assumes that each feature in the selected subset cannot have the same degree of relevance. Instead, it assigns a weight to each feature usually within the range of 0 and 1. Accordingly, feature weighting can be treated as a general version of feature selection. From the theoretical perspective of K-means defined by Eq. (1), the generalized criterion of feature weighting approaches proposed for K-means can be defined as follows:

$$W(Z, C, W) = \sum_{k=1}^{K} \sum_{z_p \in C_k} \sum_{f_i \in F} w_{f_i} (z_{pi} - c_{ki})^2 \tag{3}$$

where w_{f_i} is the weight of feature f_i in subset F, and z_{pi} represents the *ith* feature of the *pth* instance in the data.

A variety of feature weighting approaches have been developed for K-means since the 1980s. In this study, we try to discuss the most popular and innovative ones. Synthesized Clustering (SYNCLUS) (DeSarbo and Cron 1988) is treated as the first weighting approach utilized in K-means. In addition to weights representing the degree of relevance of each feature, SYNCLUS also considers weights of feature groups. Like as K-means, a user is requested to provide the number of clusters and the input data in the first step. Although it is the beginning work in this field, it may be computationally expensive and its performance heavily depends on the parameter values. However, researchers have tried to extend its performance using different techniques such as the Polak-Rebiere optimization scheme (Polak 1969) and a general linear transformation of features. Similar to SYNCLUS, FG-K-means considers two types of weights which are respectively used for features and feature groups. Nevertheless, unlike SYNCLUS, the weights of feature groups do not need to be predefined by a user. In FG-K-means, the objective function is designed by adopting both types of weights as two distinct components into the criterion of K-means defined by Eq. (1). The balance between the components and the K-means criterion is preserved by two control parameters. However, it is not easy to determine which values of the control parameters are better suited for the dataset.

The convex K-means (CK-means) approach (Modha and Spangler 2003) integrated an adaptive weighting scheme in K-means by considering different feature spaces according to the datasets. CK-means attempts to iteratively determine the optimal weights of a feature set with the goal of minimizing the average within-cluster distance. It has achieved promising results; however, it does not guarantee the optimal weights and may converge to local minima due to the gradient descent search. Another different weighting scheme for K-means (FWSA) was designed as an optimization problem by Tsai and Chio (2008). In this scheme, a feature weight is updated by simultaneously minimizing the within-cluster distance and maximizing the between-cluster separation. Unlike other weighting approaches proposed for K-means, FWSA does not require a predefined parameter.

Another feature weighting approach proposed for K-means, attribute weighting clustering (AWK) (Chan et al. 2004) assumes that each feature may have different weights of relevance at different clusters. AWK aims to minimize the sum of weighted distances within the clusters. Huang et al. (2005) proposed the weighted K-means (WK-means) which is very similar to AWK. However, there exists only a single weight for each feature in WK-means. WK-means was also integrated into fuzzy clustering. This variant, unlike the latter one, allows different weights of relevance for different clusters. WK-means and its variant have also achieved promising results but are highly dependent on a predefined control parameter which keeps the weights at a reasonable level. Moreover, there exists no such clear strategy to predefine this control parameter. Ji et al. (2013) introduced an improved version of the k-prototypes clustering approach (IK-P) which aims to minimize the WK-means criterion. IK-P is able to handle both numerical and categorical features by applying two distance criteria, respectively: (1) the Manhattan distance and (2) the frequency-based distance. IK-P produced promising results compared to k-prototype and KL-FCM-GM (Chatzis 2011). Another extension of WK-means, the intelligent WK-means (iMWK-means) was proposed by De Amorim and Mirkin (2016). Different from previous variants

of Wk-means, iMWK-means can automatically detect the number of clusters or the cluster structure in the data with the help of the Minkowski Score, while determining feature weights at the same time.

The entropy-weighting K-means (EW-KM) approach (Jing et al. 2007) maximizes the negative entropy as well as minimizes the within-cluster distance to deal with problems related to identifying such clusters using only a few dimensions. According to the results reported in Jing et al. (2007), EW-KM outperformed a number of clustering approaches. However, it is difficult to predefine the control parameter between the within-cluster distance and the negative entropy.

Inspired by K-means and locally adaptive clustering algorithms (Domeniconi et al. 2007), Parvin et al. (2013, 2015) introduced weighted locally clustering algorithms (WLAC and FWLAC). Like K-means, WLAC and FWLAC first select k well-scattered points as initial centroids and then attempt to improve the initial centroids, feature weights and cluster weights by exploring the subspace near the centroids. However, their performance is very sensitive to the control parameters.

3.2.2 Feature selection for model-based clustering

Model-based clustering generally considers all features in the modelling. However, this increases the complexity of the built model in many situations. As discussed in previous subsections, some features may not be beneficial and even may be detrimental to the clustering process. Even though all features are assumed to include clustering information, it may be problematic due to the general term, referred to as the curse of dimensionality. Especially in recent years, lots of model-based feature selection approaches have been developed for clustering. Due to the page limit, we will focus only on reviewing the most typical ones. For more information concerning this issue, please see a recent discussion in (Fop et al. 2018).

Vaithyanathan and Dom (1999) introduced a Bayesian approach which finds the best model through the marginal and integrated likelihood. This approach selects features by dividing features into clusters and then eliminating the feature clusters with irrelevant and redundant features. The feature grouping performance highly affects the feature selection process. Dy and Brodley (2004) introduced a wrapper framework (FSSM) to perform feature selection in clustering. In FSSM, the sequential forward search is used for feature selection and expectation-maximization is used as the clustering algorithm; maximum likelihood and scatter trace criteria are individually used for the feature selection procedure. To recover the bias problems of maximum likelihood and scatter trace criteria, normalization processes are applied. According to a variety of experiments, feature selection was proved to have a vital role in clustering tasks, especially on noisy problems. Despite performing normalization, the bias problems of the feature selection procedure in FSSM could not be entirely avoided. Tadesse et al. (2005) introduced a Bayesian approach that uses latent variables (features) to identify discriminating features. In this approach, a reversible jump Markow Chain method is used to create or delete clusters. Kim et al. (2006) introduced a similar approach by redesigning clustering according to an infinite mixture of distributions via Dirichlet mixtures. These approaches do not consider the correlation between relevant and irrelevant features for the clustering.

Raftery and Dean (2006) introduced a feature selection approach for model-based clustering. Features are considered as two nested sets, one of which comprises of both relevant and irrelevant features that carry cluster information, whereas the remaining set comprising

of redundant features is conditionally independent of the other one. Feature selection is carried out by comparing these two nested sets over Bayes factors on greedy search mechanism. However, this does not guarantee to get the optimal feature subset. Moreover, this approach can be time-consuming due to the comparisons of nested sets via Bayesian rules. In (Tadesse et al. 2005; Kim et al. 2006; Raftery and Dean 2006), it is assumed that irrelevant features depend on relevant feature through linear regression, i.e., irrelevant and relevant features are not fully independent. This assumption in the regression requires additional parameters but does not achieve a significant increase in the clustering performance. Maugis et al. (2005) proposed an extended version of Raftery and Dean (2006) by allowing some irrelevant features to be independent of relevant features though greedy feature selection in the regression. By this way, it is aimed to increase the clustering performance. However, the general schedule of the approach becomes even more complicated.

Zeng and Cheung (2006) developed a variant of the rival penalized expectation-maximization algorithm (RPEM) with feature weighting (FW-RPEM) to simultaneously perform clustering and feature selection. In this approach, redundant features are eliminated using the Markov blanket filter. It performed better than RPEM, but the experiments were conducted on only three datasets.

3.2.3 Feature selection for evolutionary clustering

Thanks to their effective global search abilities, EC techniques have been applied to address clustering as well as a variety of other fields. According to a number of studies in the literature (Hancer et al. 2012, 2013; Ozturk et al. 2015), EC-based clustering approaches can achieve better clustering performance than well-known clustering approaches, such as K-means and FCM. However, irrelevant and redundant features in the data may also deteriorate the performance of EC-based clustering approaches. Another significant issue associated with clustering is that the cluster structure of today's data (e.g., the number of clusters) is mostly unknown. Moreover, it is not easy to manually determine the cluster structure or the number of clusters in the data. To cover these issues, the generalized criterion for EC techniques in terms of clustering and feature selection can be defined as follows (Sheng et al. 2005).

$$Fitness = VI \times D_{weight} \times K_{weight} \tag{4}$$

where VI represents internal validity index which measures the goodness of generated clusters, D_{weight} represents a fractional or Gaussian function which measures the ratio of the selected subset size to the number of all features, and K_{weight} represents a type of Gaussian function which measures the distribution of the determined cluster number. For more information, please see (Hancer 2018).

The feature selection approaches for evolutionary clustering investigated in this paper will be further discussed in terms of the number of objectives, i.e. single-objective and multi-objective approaches.

(1) Single-objective approaches Sheng et al. (2005) proposed a variable-length GA, named as NMA-CFS, which selects features while performing clustering at the same time. In this approach, the scatter trace criterion is used to evaluate the goodness of clusters. As the first study in EC-based simultaneous clustering and feature selection, NMA-CFS obtained promising results, but the datasets used in experiments only included a small number of features and a relatively low number of clusters. In other words, it is not clear how well NMA-CFS would perform while the number of features and clusters increase.

Moreover, not only the selected features, all features are used during the computation of the scatter trace since this criterion tends to be biased over a large number of features. Das et al. (2016) introduced a GA-based simultaneous clustering and feature selection approach. To evaluate the goodness of solutions, the DBI index (Davies and Bouldin 1979) and the number of features are used in a weighted manner. In order to prevent bias during the computation of the DBI index, all features, not only selected ones, are used. Although the experimental results showed the superiority of the GA-based approach against FCM, the performance was only evaluated through the best solutions, not considering the stochasticity of the algorithm that requires the evaluations to be done based on the average of multiple solutions.

Sarvari et al. (2010) redesigned the concept of NMA-CFS using the harmony search (HS) algorithm and the centroid-based encoding scheme. Cobos et al. (2010) introduced a hybridized K-means and HS (IHSK) approach to performing simultaneous clustering and feature selection. In IHSK, a local add-remove operator is sequentially performed to select feature subsets and thereby the fitness function does not include D_{weight} component. If it is possible to reach better fitness value than current fitness value by adding (removing) any feature into (from) the feature subset, this feature will be selected (removed). Accordingly, only the selected features are used during the calculation of cluster validity criterion. Another local search module in IHSK, K-means is carried out for a few iterations to decrease the adverse effects of initial conditions. From the results, it can be inferred that IHSK could achieve better performance than a variety of approaches such as NMA-CFS and K-means. However, it was only tested on three real-world datasets.

Swetha and Devi (2012) introduced a two-stage particle swarm optimization (PSO) based approach for clustering which first performs feature selection and then applies clustering on the selected feature subset. Javani et al. (2011) introduced another PSO-based simultaneous clustering and feature selection approach. In this approach, a new kernelized validity index is defined and K-means is probabilistically called to decrease the adverse effects of initial conditions during the evolutionary process. However, this approach was not tested on high dimensional datasets. Prakash et al. (2015) introduced a hybridized K-means and binary PSO approach (BPSO-X) for simultaneous clustering and feature selection. In BPSO-X, each solution is encoded via 0 and 1 representing a possible feature subset. To evaluate the quality of a solution, BPSO-X first selects the possible feature subset and then performs K-means on the reduced dataset to obtain clusters. After K-means, the obtained clusters are evaluated using the Silhouette index (Rousseeuw 1987). Like as IHSK, BPSO-X also considers the selected feature subset to construct clusters. Although a number of experiments were conducted in (Prakash and Singh 2015) on a variety of datasets by comparing BPSO-X with BPSO and GA, it is not possible to make a consistent analysis due to the following drawbacks. First, traditional clustering approaches such as K-means and FCM were not used in comparisons. Second, only the results of the best solutions were reported in terms of external validity indexes.

Lensen et al. (2016) developed PSO-based approaches using partition-based and centroid-based representation schemes for simultaneous clustering and feature selection. Inspired by NMA-CFS, PSO-based approaches use a type of the scatter trace criterion and the same feature weighting scheme (K_{weight}). On the other hand, different from NMA-CFS, PSO-based approaches use a specified logarithmic function as the component of K_{weight} to automatically identify the cluster structure. According to the results, the partition-based representation scheme performed well compared to the centroid-based encoding scheme when K was fixed. However, the partition-based encoding scheme may be computationally intensive for datasets with a large number of instances. Lensen et al. (2017) also developed

a multiple-stage PSO-based approach. In the first stage, the Silhouette method (Rousseeuw 1987) is used to obtain an initial number of clusters (K_{est}). In the second stage, the simultaneous clustering and feature selection process is carried out using a partition-based PSO approach. In particular, the K_{est} value obtained from the previous stage is used in this stage as the mean parameter to build a Gaussian weighting function to determine the appropriate number of clusters. Moreover, a type of Gaussian function is also designed to select a feature subset from all available features. In the last stage, the partition-based encoding scheme is converted to a centroid-based encoding scheme and then the centroids are fine-tuned using another PSO search process. It is observed from the results that it generally performed better than the previous PSO-based approaches. However, it could not properly work in real-world datasets in terms of detecting the optimal number of clusters.

Hancer (2018) introduced a DE-based approach to simultaneously perform clustering and feature selection. In this approach, a similarity scheme based discrete DE variant is used to evolve solutions, and a new objective function is improved by adopting a type of Gaussian feature weighting function to the Turi's VI index (Turi 2001). According to the results, it generally performed better than traditional clustering approaches such as K-means and FCM despite the lack of information concerning the number of clusters. Moreover, it also achieved better performance than the previously described PSO-based approaches (Lensen et al. 2016, 2017) in terms of detecting the optimal number of features. However, both these DE-based and PSO-based approaches use all features during the computation of cluster validity criteria to prevent feature selection leading a bias towards a small number of clusters.

(2) Multi-objective approaches Simultaneous clustering and feature selection task has also been treated as a multi-objective problem. Kim et al. (2002) introduced a multi-objective algorithm (ELSA) using an evolutionary local selection algorithm and K-means. To evaluate the quality of clusters, a type of the within-cluster distance and the between-cluster separation are individually used as the objectives in the ELSA algorithm. However, these objectives require data-dependent constants which need to be specified according to the dataset.

Dutta (2012, 2013) proposed multi-objective GA-based simultaneous clustering and feature selection approaches. In contrast to previous ones, these approaches assume that K is fixed. A type of within-cluster distance and the between-cluster separation are employed as objectives to evaluate the goodness of solutions. According to the results, it performed better than K-means and multi-objective clustering algorithms. However, it is not possible to make a consistent analysis since the performance analysis was only made through internal indexes.

Saha et al. (2014, 2015) introduced multi-objective simulating annealing based simultaneous clustering and feature selection approaches. In contrast to previous multi-objective approaches, it considers more than two objectives which are the Sym index, the XB index, the number of features and the adjusted Rand index (for the supervised version). During the calculation of validity criteria, only the selected features represented in individuals are used. For the selection of the best solution on the final Pareto front, the class labels for a small number of instances in the dataset is assumed to be known by a user, referred to as test instances. As each solution on the final Pareto front provides a set of clusters, a cluster label is assigned to each test sample. Thereafter, the Minkowski score is computed for each solution on the final Pareto front using the cluster labels and the class labels. The solution which has the best score is chosen as the final solution. Although they achieved promising results against a variety of clustering approaches such as VGAPS and K-means, they only examined the algorithms on datasets with 2 or 3 clusters.

In another study, Saha et al. (2018) introduced a multi-objective DE-based approach (Auto-MODE) to identify microRNAs. In contrast to previous multi-objective approaches, Auto-MODE uses a weighted scheme like as the approaches proposed for K-means to select the optimal feature subset. In Auto-MODE, the weighted versions of the Xie-Beni index (Xie and Beni 1991) and the I index (Bandyopadhyay and Saha 2012) are used as the objectives. Thus, the selected feature subset size does not need to be considered as an objective. According to a number of experiments comparing with various clustering approaches, Auto-MODE achieved good clustering quality in miRNAs problems. However, the cluster quality was only analyzed through internal indexes.

Prakash and Singh (2019) introduced a genetically inspired multi-objective binary gravitational search based simultaneous clustering and feature selection approach (IMBGSAFS). In IMBGSAFS, the Silhouette index (Rousseeuw 1987) and the feature subset size are the objectives used to search the possible solution space, and an external archive is used to build the non-dominated solutions set. For clustering, K-means is applied on the reduced dataset using the selected features in a solution. After the evolutionary process, the solution which has the best F-measure score is selected as the final solution. According to the results, IMBGSAFS performed better than elitist non-dominated sorting GA (NSGAII (Deb et al. 2002)). To achieve better results using NSGAII, improved NSGAII variants (Lee et al. 2017; Song and Chen 2018; Gao et al. 2019) could be redesigned and employed for simultaneous clustering and feature selection.

Hancer (2020) introduced a variable-string length based multi-objective DE-based approach (MODE-CFS) for simultaneous clustering and feature selection. To evaluate the goodness of solutions, MODE-CFS simultatenously maximizes the Silhouette index (Rousseeuw 1987) and the number of selected features while minimizing the WB index. The overall goal of maximizing the number of features is to prevent overlapping problems during the distance calculations between instances for the evaluation of validity indexes. For each solution, a mutant solution is generated using a specified two-case mutation scheme, and then the current solution and its mutant are added to a union set. Finally, a non-dominated sorting scheme is applied to find fitter solutions for the next generation. At the end of the evolutionary process, a single solution is selected using a semi-supervised selection scheme inspired by Saha et al. (2014, 2015). According to the results, it outperformed a variety of conventional and recently introduced multi-objective clustering approaches.

3.3 Embedded approaches

Embedded approaches carry out the process of feature subset selection as a part of the learning process. Therefore, embedded approaches are computationally less expensive than wrapper approaches. For embedded approaches, sparse learning algorithms have a very crucial role due to their performance and interpretability. Sparse learning algorithms aim to seek a trade-off between some goodness measure and sparsity of the result. For example, in a sparse learning clustering task, the cluster quality or some other typical measure of performance is not the only concern: we also wish to be able to clarify what the clustering approach means to a non-expert user. Therefore, if a biomedical data is taken into consideration for the clustering process, we should provide a clustering approach that achieves high-quality clusters, but also represent the biomedical data with a few features.

The idea of applying sparse learning in clustering has attracted researchers' attention in recent years. Typically, these approaches first find the cluster labels using a clustering algorithm and then transform the unsupervised feature selection into a supervised context

through the generated cluster labels. In particular, it is aimed to preserve the manifold structure constrained from the whole feature space. The flowchart of an embedded sparse learning approach is presented in Fig. 3.

Treated as the earliest sparse learning feature selection approach proposed for clustering, multi-cluster feature selection (MCFS) (2010) has three stages. In the first stage, spectral analysis is used to explore the intrinsic structure, i.e., to measure the correlation between features. In the second stage, a regression model with L1-regularized least-square is performed to quantify the importance of the features. In the last stage, a specified number of features with the highest values of coefficients obtained through the previous stage is selected. Although MCFS has proven to be a well-designed approach to address feature selection for clustering, it is computationally expensive. To improve the efficiency of MCFS, Wang and Shen (2016) integrated the locality sensitive hasing forest (Bawa et al. 2005) in MCFS. Another variant of MCFS (MCFS-SDS) (Liu and Liu 2012) was improved by simultaneously evaluating L1-regularized least-square and L2-regularized least-square through smooth distributed score. MCFS-SDS is therefore more efficient than the pure MCFS especially for the small number of clusters.

Another earlier sparse learning feature selection approach proposed for clustering, MRSF (Zhao et al. 2010) considers the feature selection task as a multi-output regression problem, and applies $L_{2,1}$-regularized least-square instead of L1-regularized least-square. Inspired by MRSF, Yang et al. (2011) introduced an unsupervised feature selection approach (UDFS) which utilizes the local discriminative information contained in the feature correlations and scatter matrices. By using local information, UDFS well preserves the cluster structure. For the selection of feature subset, $L_{2,1}$-norm regularizer with some additional constraints is used. Including similar properties with UDFS, nonnegative discriminative feature selection (NDFS) (Li et al. 2012) simultaneously performs spectral clustering and feature selection. The cluster labels are obtained by spectral clustering to carry out feature selection. A nonnegative constraint is adapted to the class labels to get more accurate cluster labels. For the procedure of feature selection, L2-regularized least-square is used.

To deal with noise in the data, Qian and Zhai (2013) introduced a spectral learning feature selection algorithm for clustering, called robust unsupervised feature selection (RUFS). Unlike MCFS, UDFS and NDFS, the cluster labels are determined using local learning regularized nonnegative matrix factorization. For feature selection, a robust joint $L_{2,1}$-norm is used while determining the cluster labels. Following similar concept with RUFS, robust unsupervised feature selection via matrix factorization (RUFSM) (Du et al. 2017) simultaneously performs discriminative feature selection and robust clustering through $L_{2,1}$-norm to select features. Unlike RUFS, RUFSM uses the cluster centroids rather than the cluster labels.

As seen in Fig. 3, a conventional embedded sparse learning feature selection approach like MCFS, NDFS and RUFS requires the cluster labels generated by a clustering algorithm

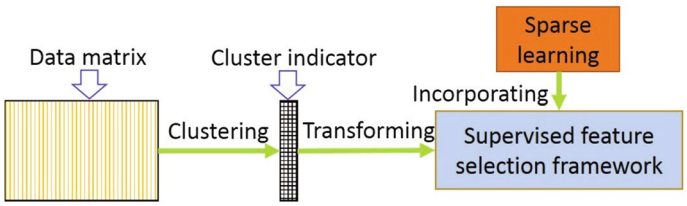

Fig. 3 Working scheme of a typical embedded sparse learning feature selection approach (Wang et al. 2017)

to transform unsupervised feature selection into supervised feature selection. However, this transformation generally causes non-optimal feature subsets. To address this issue, Wang et al. (2017) proposed a novel embedded approach called embedded unsupervised feature selection (EUFS), which directly embed feature selection into a clustering algorithm without the transformation. EUFS applies K-means by minimizing reconstruct error to obtain the cluster labels and selects features. However, EUFS tends to select large-magnitude features that may be non-discriminative. To address this issue, Zhu and Yang (2018) proposed discriminative embedded unsupervised feature selection approach (DEUFS) which obtains the cluster labels by maximizing the heterogeneity between clusters to model the cluster structure of data. The general flowchart of EUFS and DEUFS approaches is presented in Fig. 4.

3.4 Hybrid approaches

As mentioned in Sect. 2.1, filters generally complete the selection process of a feature subset in a shorter time than wrappers. On the other hand, filters cannot enhance the performance of a learning algorithm as well as wrappers. This is because wrappers synchronously work with a learning algorithm. To bring the advantages of both wrappers and filters, i.e., to improve the computational efficiency of the selection process and the performance of the further learning processes, hybridized wrapper-filter approaches have become very common in the literature (Hancer 2019).

Dash and Liu (2000) introduced a hybridized two-way feature selection approach for clustering. In the first stage, features are ranked based on the entropy measure of information theory. To rank features, each feature is removed from the available feature set and its degradation effect is measured by calculating the entropy of the remaining features set. In the second stage, the top-ranked feature is selected from the ranked feature set, and then K-means is applied on the selected feature subset for each iteration. This procedure is repeated until reaching the highest value of the scatter trace criterion for the selected feature subset.

Another hybridized two-way feature selection approach (Yun Li et al. 2006) first tries to eliminate redundant features using an entropy-based measure and the fuzzy evaluation index (FFIE) (Pal et al. 2000). After the elimination stage of redundant features, it tries to select an optimal feature subset among the reduced feature set using FCM and the scatter trace criterion. However, there is no comprehensive experimental analysis performed to evaluate the performance of the hybridized wrapper-filter approach.

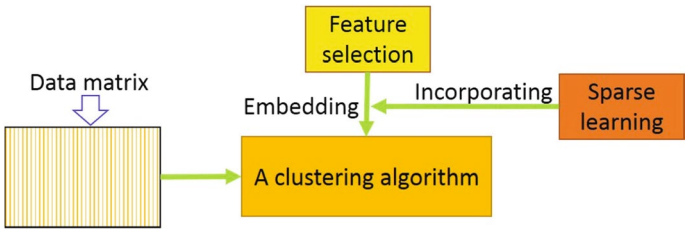

Fig. 4 Working scheme of a embedded sparse learning feature selection approach without transformation (Wang et al. 2017)

Solorio-Fernández et al. (2016) proposed a hybridized two-stage approach based on the Laplacian Score (He et al. 2005) and the Calinski–Harabasz index (Calinski and Harabasz 1974). In the first stage, features in the data are ranked using the Laplacian Score. In the second stage, K-means is applied on the selected feature subsets obtained by forward and backward strategies to generate cluster sets, and then the generated cluster sets are evaluated using a modified version of Calinkski-Harasz called the Weighted Normalized Calinski–Harabasz index (WNCH). Among possible feature subsets, the one with the highest WNCH value is chosen as the optimal feature subset.

Jaskhi et al. (2009) introduced a wrapper-filter approach for document clustering. The approach first randomly labels the document with random labels. It then selects a feature subset using a Bayesian model. In the final step, K-means is applied on the selected feature subset. This procedure is repeated until the stopping criterion is met. Another wrapper-filter approach (Kim et al. 2006) tries to find a number of features by applying the least-square estimation (LSE) criterion (Mao 2005) and then applies a modified version of the expectation-maximization algorithm to select the optimal feature subset that maximally improves the performance of clustering.

Hruschka et al. (2005) introduced a wrapper-filter feature selection approach (BFK), which combines K-means and Bayesian network. In contrast to the previous wrapper-filter approaches, BFK first carries out the wrapper stage by applying K-means with a range of clusters. The cluster structure with the highest value of the Silhouette index (Rousseeuw 1987) is chosen for the further stage. In other words, the cluster structure is automatically evolved without the requirement of k number of clusters using a traditional approach. In the filter stage, the feature subset is chosen with the help of a Bayesian network, which considers each cluster and feature as a class and a node, respectively. Although it can be treated as a novel wrapper-filter approach when compared to the previous wrapper-filter approaches, its performance heavily depends on the Silhouette index. For instance, if the cluster structure is not correctly determined using the Silhouette index, the further filter process for feature selection would be adversely affected. For some more wrapper-filter approaches proposed for clustering, please see (Yang et al. 2011; Zhang et al. 2012; Fan et al. 2013).

4 Further discussions

4.1 Summary of studies

The main focus of this section is to provide a general summarization and discussions concerning feature selection for clustering. From feature selection approaches proposed for clustering, it can be found that the main goal of feature selection is to increase the clustering quality, reduce the storage requirements and increase the time-computational efficiency. However, there are still various limitations in existing work, such as the need of predefining the number of clusters, the number of features, or data with a small number of clusters or insufficient analysis on high dimensional datasets. How to apply feature selection for clustering is still an open issue and has not been fully addressed yet. Should the feature selection process be before or after the clustering process? Alternatively, should feature selection and clustering be simultaneously or synchronously performed? Which way we will use for clustering depends on whether the cluster structure is preserved or not.

In case of feature selection being applied before clustering, the feature subset is first selected from the whole data and then apply clustering on the selected feature space. In

other words, the feature selection process is independent of the clustering process. This way is considered in filter feature selection approaches in this work. Filter approaches can be easily applied to both supervised and unsupervised learning tasks. These approaches aim to build an affinity/similarity matrix and then a score is obtained for each feature. Although these approaches are simple and straightforward, these approaches cannot deal with feature redundancy, since they repeatedly tend to select highly correlated features during the feature selection process.

In the second way, feature selection is performed synchronously with clustering. Thus, the selected feature subset is dependent on the clustering algorithm. This way is often referred to as wrappers in the terminology of feature selection. In this work, wrappers are investigated in three groups: feature selection for K-means, feature selection for model-based and feature selection for EC-based clustering algorithms. As K-means is considerably treated as one of the most popular clustering algorithms among researchers, a significant number of efforts have been put forward to enhance the clustering performance of K-means. One solution to this issue is to integrate a feature weighting scheme into the K-means algorithm. By this way, it is aimed to increase the clustering performance of K-means by removing irrelevant, redundant and weakly redundant features in the data. A variety of feature weighting approaches have been proposed for K-means in the literature. According to the results, they significantly improved the performance of the K-means algorithm. However, they still suffer from some challenges: initial conditions, the dependence of predefined parameters and premature convergence to a local optimum. Moreover, we observe that a large number of studies on feature selection for K-means did not introduce a comprehensive experimental study. For instance, in Chan et al. (2004), Tsai and Chiu (2008), only two or three real-world benchmarks have been used for the experimental studies. For another example, Amorim et al. (2013) introduced a comprehensive comparative study of feature selection approaches for K-means in a variety of benchmark datasets. However, the results of the pure K-means algorithm were not presented in the experiments. It lacks of a consistent analysis on how feature weighting affects the performance of K-means.

Another category of wrappers, in contrast to feature selection approaches for K-means, feature selection approaches for model-based clustering do not require extra predefined parameters. On the other hand, they can only detect gaussian distributed clusters, i.e., they cannot properly work for the detection of unshaped or uniformly distributed clusters. Accordingly, it may not be possible to obtain optimal clustering performance on real-world datasets which are mostly in the form of unshaped distribution.

The last category of wrappers is the feature selection approaches proposed for evolutionary clustering. We investigate feature selection approaches to evolutionary clustering according to the number of objectives. As for single-objective approaches, the most common way is to allow feature selection and clustering processes simultaneously. In detail, such approaches try to optimize both the clustering performance and the feature subset size of the possible selected feature subset. Although these approaches are able to find well-separated clusters in the data, all features in the data, not only the selected ones, are considered to measure the clustering performance due to the bias problems towards a small number of clusters. In other words, the quality of the clusters is not evaluated correctly through the selected feature subset. Another way of single-objective feature selection approaches is to apply local add-remove operators (to add and/or remove features) during the evolutionary process. This way is not very common like the previous ways, but only the selected features are used to evaluate the clustering performance. However, the approaches using local add-remove operators could not reach promising clustering performance. When considering

multi-objective approaches, there exists no such a challenge that leads bias problems between clustering and feature selection like most of the single-objective approaches, i.e., the clustering performance of a selected feature set can be evaluated without any requirements. However, which solution is chosen from the Pareto front is a challenging task. To address this issue, a semi-supervised selection scheme is applied on a small number of solutions. Moreover, the studies on multi-objective approaches have recently come into consideration and are still at the beginning level compared to single-objective approaches.

In case of feature selection being applied synchronously with clustering, embedded approaches have also been playing an important role, especially in recent years. The two general working schemes of embedded approaches are given as follows. In the first scheme, a clustering algorithm is first applied to obtain the pseudo cluster labels and then unsupervised feature selection is transformed into supervised concept using the generated labels. In other words, the cluster labels are treated as kind of class labels in order to perform feature selection. If the generated clusters (i.e., the cluster labels) are not real or optimal ones, it is hard for the feature selection process to achieve good results. In the second scheme, feature selection is directly adopted into clustering without the requirement of transformation. The second scheme tends to work well in datasets with large magnitude features compared to the first scheme. Both schemes are built using the principles of sparse-learning approaches. Although embedded sparse-learning based feature selection approaches have produced promising results, there are still some challenges concerning these approaches: complex matrix operations and intensive computational cost. Moreover, there are only a few feature selection approaches wrapped around the scheme in the literature. Notice that, some sparse-learning based feature selection approaches (e.g. SPEC) can be evaluated in the category of filters.

The last way considered in this study is to hybridize wrappers and filters in a framework. Although hybrid approaches are believed to improve the effectiveness and efficiency of a learning algorithm, they are not very common in the literature compared to wrappers and filters. The most commonly followed technique to build a hybridized wrapper-filter model is the two-way hybrid approaches, which first employ a filter approach to reduce the dimensionality and then apply a wrapper approach on the reduced feature set to find fitter feature subsets. According to a variety of studies in the literature, two-way approaches have reached promising results for clustering. However, there is no interaction between the stages in two-way approaches, i.e., both stages carry out their process independent of each other. Therefore, there is a need for hybridized wrapper-filter approaches that synchronously combine wrapper and filter stages. One solution to this issue is to integrate a local search filter (wrapper) mechanism into the wrapper (filter) framework. There exist some such wrapper-filter approaches that use local search mechanisms to combine wrapper and filter stages in the literature proposed for classification (Hancer 2019; Butler-Yeoman et al. 2015), but developing such wrapper-filter approaches for clustering is still an open issue.

4.2 Future trends

There have been many feature selection algorithms proposed for clustering, which achieve some success but also have limitations. Feature selection itself is a challenging task and unsupervised feature selection is even harder due to the lack of target labels of instances. There are still many space for future research in this area, such as the scalability, speed, and evaluation measures. Along with the tremendous growth of dataset sizes, the scalability of current clustering algorithms may become a big issue. Feature

selection approaches proposed for clustering faces a similar problem. However, most clustering approaches require to keep all features in the memory to observe any evolving in the cluster structure. If any difference is observed in the data structure, the clustering process should be started again. Furthermore, the scalability of feature selection approaches is a big problem. Feature selection approaches usually require a sufficient number of data instances to obtain good learning performance. To address this issue, some approaches memorize only important instances or prototypes, i.e., the mean of each cluster. Therefore, the scalability of clustering and feature selection approaches should be given more attention to keeping pace with the growth and fast streaming of the data.

Computational time is often a big issue in most data mining tasks, even for datasets with medium sizes. Feature selection has a large search space, which grows exponentially with the number of original features in the dataset. Feature selection for clustering, especially wrapper methods, is very expensive due mainly to that each evaluation requires to perform a clustering process to test the goodness of the selected feature subset. This process becomes even slower when the dataset size is big. Although feature selection for clustering can reduce the computation time for future applications, the process of selecting features itself needs to be speeded up. This could be achieved by developing an efficient search mechanism for feature selection, or a two-stage method to quickly filter out useless features, or fast evaluation measures.

Evaluation measures are an important factor that determines the speed and the final performance of feature selection and clustering. Clustering as an unsupervised method is harder than classification, where the training set has ground truth class labels for the learning algorithm to learn from, and the performance measures are easy to determine, e.g., accuracy or F-measures. There are many clustering evaluation measures available, e.g. various within or between cluster distances, compactness, separation and connectedness. There are also internal measures and external measures. However, different evaluation measures often show different level of goodness for the same set of clustering results, and suggest different methods as the best algorithms. Therefore, it is hard for users to make a decision. It is a very challenging task to develop powerful and comprehensive evaluation measures, but will greatly benefit the community.

5 Conclusions

In this paper, we introduced a comprehensive survey on feature selection approaches to clustering. To better summarize the profile of this field, we have introduced a simple categorization in terms of the evaluation, i.e., filter methods, wrapper methods, embedded methods and hybrid methods. Moreover, wrappers are further categorized as K-means, model-based and evolutionary clustering based algorithms. Based on this categorization, we have discussed the corresponding works in detail, including their key ideas, results and limitations. Another contribution of this paper is the thorough discussions of feature selection approaches to evolutionary clustering. To the best of our knowledge, there exists no such work that focuses on feature selection approaches designed for evolutionary clustering. We also discussed possible future trends and challenges in this area, such as the scalability, the speed of the algorithms, and powerful evaluation measures in feature selection for clustering.

References

Alelyani S, Tang J, Liu H (2013) Feature selection for clustering: a review. In: Aggarwal CC, Reddy CK (eds) Data clustering: algorithms and applications

Aloise D, Deshpande A, Hansen P, Popat P (2009) Np-hardness of Euclidean sum-of-squares clustering. Mach Learn 75(2):245–248

Amini S, Homayouni S, Safari A, Darvishsefat AA (2018) Object-based classification of hyperspectral data using random forest algorithm. Geo Spat Inf Sci 21(2):127–138

Ankerst M, Breunig MM, Kriegel HP, Sander J (1999) Optics: ordering points to identify the clustering structure. In: Proceedings of the ACM SIGMOD international conference on management of data, SIGMOD'99. ACM, New York, NY, USA, pp 49–60

Awad MM (2018) Forest mapping: a comparison between hyperspectral and multispectral images and technologies. J For Res 29(5):1395–1405

Bandyopadhyay S, Saha S (2012) Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications. Springer, Berlin

Bawa M, Condie T, Ganesan P (2005) LSH forest: self-tuning indexes for similarity search. In: Proceedings of the 14th international conference on world wide web, WWW'05. ACM, New York, NY, USA, pp 651–660

Bezdek JC, Ehrlich R, Full W (1984) FCM: The fuzzy c-means clustering algorithm. Comput Geosci 10(2):191–203

Butler-Yeoman T, Xue B, Zhang M (2015) Particle swarm optimisation for feature selection: a hybrid filter-wrapper approach. In: IEEE congress on evolutionary computation (CEC), pp 2428–2435

Cai D, Zhang C, He X (2010) Unsupervised feature selection for multi-cluster data. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'10. ACM, New York, NY, USA, pp 333–342

Calinski R, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3(1):1–27

Chakraborty S, Das S (2018) Simultaneous variable weighting and determining the number of clusters—a weighted Gaussian means algorithm. Stat Probab Lett 137:148–156

Chan EY, Ching WK, Ng MK, Huang JZ (2004) An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recognit 37(5):943–952

Chatzis SP (2011) A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. Expert Syst Appl 38(7):8684–8689

Cheung Y, Zeng H (2006) Feature weighted rival penalized em for gaussian mixture clustering: automatic feature and model selections in a single paradigm. Int Conf Comput Intell Secur 1:633–638

Cobos C, Leon E, Mendoza M (2010) A harmony search algorithm for clustering with feature selection. Rev Fac Ing Univ Antioq 55:153–164

Das S, Chaudhuri S, Ghatak S, Das AK (2016) Simultaneous feature selection and cluster analysis using genetic algorithm. In: International conference on information technology (ICIT2016), pp 288–293

Dash M, Liu H (1999) Handling large unsupervised data via dimensionality reduction. In: SIGMOD research issues in data mining and knowledge discovery (DMKD-99) workshop

Dash M, Liu H (2000) Feature selection for clustering. In: Terano T, Liu H, Chen ALP (eds) Knowledge discovery and data mining. Current issues and new applications, pp 110–121

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(2):224–227

de Amorim RC (2016) A survey on feature weighting based k-means algorithms. J Classif 33(2):210–242

DeSarbo WS, Cron WL (1988) A maximum likelihood methodology for clusterwise linear regression. J Classif 5(2):249–282

Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197

Dheeru D, Karra Taniskidou E (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml

Domeniconi C, Papadopoulos D, Gunopulos D, Ma S (2004) Subspace clustering of high dimensional data. In: Siam international conference on data mining

Domeniconi C, Gunopulos D, Ma S, Yan B, Al-Razgan M, Papadopoulos D (2007) Locally adaptive metrics for clustering high dimensional data. Data Min Knowl Discov 14(1):63–97

Dorigo M, Di Caro G (1999) Ant colony optimization: a new meta-heuristic. Proc Congr Evol Compu 2:1470–1477

Du S, Ma Y, Li S, Ma Y (2017) Robust unsupervised feature selection via matrix factorization. Neurocomputing 241:115–127

Dutta D, Dutta P, Sil J (2012) Simultaneous feature selection and clustering for categorical features using multi objective genetic algorithm. In: 12th international conference on hybrid intelligent systems (HIS2012), pp 191–196

Dutta D, Dutta P, Sil J (2013) Simultaneous continuous feature selection and k clustering by multi objective genetic algorithm. In: 3rd IEEE international advance computing conference (IACC2013), pp 937–942

Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5:845–889

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second international conference on knowledge discovery and data mining, KDD'96, pp 226–231

Fan W, Bouguila N, Ziou D (2013) Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference. IEEE Trans Knowl Data Eng 25(7):1670–1685

Ferreira AJ, Figueiredo MA (2012) An unsupervised approach to feature discretization and selection. Pattern Recognit 45(9):3048–3060

Fop M, Murphy TB, Scrucca L (2018) Model-based clustering with sparse covariance matrices. Stat Comput 39:1–29

Gao D, Liang H, Shi G, Cao L (2019) Multi-objective optimization of carbon fiber-reinforced plastic composite bumper based on adaptive genetic algorithm. Math Problems Eng. https://doi.org/10.1155/2019/8948315

Golub GH, Reinsch C (1970) Singular value decomposition and least squares solutions. Numer Math 14(5):403–420

Grün B (2019) Model-based clustering. CRC Press, Boca Raton, pp 163–198

Guha S, Rastogi R, Shim K (1998) Cure: an efficient clustering algorithm for large databases. SIGMOD Rec 27(2):73–84

Guha S, Rastogi R, Kyuseok S (1999) Rock: a robust clustering algorithm for categorical attributes. In: 15th international conference on data engineering, 1999. Proceedings, pp 512–521

Haindl M, Somol P, Ververidis D, Kotropoulos C (2006) Feature selection based on mutual correlation. In: Carrasco Ochoa JA, Kittler J, Martínez-Trinidad JF (eds) Progress in pattern recognition, image analysis and applications. Springer, Berlin, pp 569–577

Hancer E (2019) Differential evolution for feature selection: a fuzzy wrapper-filter approach. Soft Comput 23(13):5233–5248

Hancer E (2020) A new multi-objective differential evolution approach for simultaneous clustering and feature selection. Eng Appl Artif Intell 87:103307

Hancer E, Karaboga D (2017) A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number. Swarm Evol Comput 32:49–67

Hancer E, Xue B, Zhang M (2018) Differential evolution for filter feature selection based on information theory and feature ranking. Knowl Based Syst 140:103–119

Hancer E (2018) A differential evolution approach for simultaneous clustering and feature selection. In: International conference on artificial intelligence and data processing, pp 1–7

Hancer E, Ozturk C, Karaboga D (2012) Artificial bee colony based image clustering method. In: IEEE congress on evolutionary computation (CEC2012), pp 1–5

Hancer E, Ozturk C, Karaboga D (2013) Extraction of brain tumors from MRI images with artificial bee colony based segmentation methodology. In: 8th international conference on electrical and electronics engineering (ELECO2013), pp 516–520

Hancer E, Samet R, Karaboga D (2014) A hybrid method to the reconstruction of contour lines from scanned topographic maps. In: IEEE 23rd international symposium on industrial electronics (ISIE2014), pp 930–933

He X, Cai D, Niyogi P (2005) Laplacian score for feature selection. In: Proceedings of the 18th international conference on neural information processing systems, NIPS'05. MIT Press, Cambridge, MA, USA, pp 507–514

Hinneburg A, Gabriel HH (2007) Denclue 2.0: fast clustering based on kernel density estimation. In: Shawe-Taylor MRBJ, Lavrač N (eds) Advances in intelligent data analysis VII, pp 70–80

Holland JH (1975) Adaption in natural and artificial systems. University of Michigan Press, Ann Arbor

Hruschka ER, Campello RJGB, Freitas AA, De Carvalho ACPLF (2009) A survey of evolutionary algorithms for clustering. IEEE Trans Syst Man Cybern Part C Appl Rev 39(2):133–155

Hruschka ER, Hruschka ER, Covoes TF, Ebecken NFF (2005) Feature selection for clustering problems: a hybrid algorithm that iterates between k-means and a Bayesian filter. In: Fifth international conference on hybrid intelligent systems (HIS'05), pp 1–6

Huang JZ, Ng MK, Rong H, Li Z (2005) Automated variable weighting in k-means type clustering. IEEE Trans Pattern Anal Mach Intell 27(5):657–668
Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
Jashki MA, Makki M, Bagheri E, Ghorbani AA (2009) An iterative hybrid filter-wrapper approach to feature selection for document clustering. In: Gao Y, Japkowicz N (eds) Advances in artificial intelligence. Springer, Berlin, pp 74–85
Javani M, Faez K, Aghlmandi D (2011) Clustering and feature selection via PSO algorithm. In: 2011 international symposium on artificial intelligence and signal processing (AISP), pp 71–76
Ji J, Bai T, Zhou C, Ma C, Wang Z (2013) An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neurocomputing 120:590–596
Jing L, Ng MK, Huang JZ (2007) An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans Knowl Data Eng 19(8):1026–1041
Jolliffe I (1986) Principal component analysis. Springer, Berlin
Karaboga D, Gorkemli B, Ozturk C, Karaboga N (2014) A comprehensive survey: artificial bee colony ABC algorithm and applications. Artif Intell Rev 42(1):21–57
Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of international conference on neural networks (ICNN'95), vol 4, pp 1942–1948
Kim Y, Street WN, Menczer F (2002) Evolutionary model selection in unsupervised learning. Intell Data Anal 6(6):531–556
Kim S, Tadesse MG, Vannucci M (2006) Variable selection in clustering via dirichlet process mixture models. Biometrika 93(4):877–893
Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection. MIT Press, Cambridge
Lee Y, Choi TJ, Ahn CW (2017) Multi-objective evolutionary approach to select security solutions. CAAI Trans Intell Technol 2(2):64–67
Lensen A, Xue B, Zhang M (2017) Using particle swarm optimisation and the silhouette metric to estimate the number of clusters, select features, and perform clustering. In: Squillero G, Sim K (eds) Applications of evolutionary computation. Springer, Berlin, pp 538–554
Lensen A, Xue B, Zhang M (2016) Particle swarm optimisation representations for simultaneous clustering and feature selection. In: IEEE symposium series on computational intelligence (SSCI)
Li Y, Dong M, Hua J (2008) Localized feature selection for clustering. Pattern Recognit Lett 29(1):10–18
Li Y, Lu BL, Wu ZF (2007) Hierarchical fuzzy filter method for unsupervised feature selection. J Intell Fuzzy Syst 18(2):157–169
Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2016) Feature selection: a data perspective. CoRR abs/1601.07996
Li Z, Yang Y, Liu J, Zhou X, Lu H (2012) Unsupervised feature selection using nonnegative spectral analysis. In: Proceedings of the twenty-sixth AAAI conference on artificial intelligence, AAAI'12. AAAI Press, pp 1026–1032
Liu F, Liu X (2012) Unsupervised feature selection for multi-cluster data via smooth distributed score. In: Huang DS, Gupta P, Zhang X, Premaratne P (eds) Emerging intelligent computing technology and applications. Springer, Berlin, pp 74–79
Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502
Macqueen J (1967) Some methods for classification and analysis of multivariate observations. In: 5th Berkeley symposium on mathematical statistics and probability, pp 281–297
Mao KZ (2005) Identifying critical variables of principal components for unsupervised feature selection. IEEE Trans Syst Man Cybern Part B (Cybern) 35(2):339–344
Maugis C, Celeux G, Martin-Magniette ML (2005) Variable selection for clustering with Gaussian mixture models. Biometrics 65(3):602–617
McLachlan GJ, Krishnan T (2008) The EM algorithm and extensions (Wiley series in probability and statistics), 2nd edn. Wiley, Hoboken
Miruthula P, Roopa SN (2015) Unsupervised feature selection algorithms: a survey. Int J Sci Res 4(6):688–690
Mitra P, Murthy CA, Pal SK (2002) Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Mach Intell 24(3):301–312
Modha DS, Spangler WS (2003) Feature weighting in k-means clustering. Mach Learn 52(3):217–237
Mugunthadevi K, Punitha SC, Punithavalli M, Mugunthadevi K (2011) Survey on feature selection in document clustering. Int J Comput Sci Eng 3:1240–1241
Nie F, Xiang S, Jia Y, Zhang C, Yan S (2008) Trace ratio criterion for feature selection. vol 2, pp 671–676

Ozturk C, Hancer E, Karaboga D (2015) Improved clustering criterion for image clustering with artificial bee colony algorithm. Pattern Anal Appl 18(3):587–599
Pal SK, De RK, Basak J (2000) Unsupervised feature evaluation: a neuro-fuzzy approach. IEEE Trans Neural Netw 11(2):366–376
Parsons L, Haque E, Liu H (2004) Subspace clustering for high dimensional data: a review. SIGKDD Explor Newsl 6(1):90–105
Parvin H, Beigi A, Mozayani N (2012) A clustering ensemble learning method based on the ant colony clustering algorithm. Appl Comput Math 11:286–302
Parvin H, Minaei-Bidgoli B (2013) A clustering ensemble framework based on elite selection of weighted clusters. Adv Data Anal Classif 7(2):181–208
Parvin H, Minaei-Bidgoli B (2015) A clustering ensemble framework based on selection of fuzzy weighted clusters in a locally adaptive clustering algorithm. Pattern Anal Appl 18(1):87–112
Patnaik AK, Bhuyan PK, Rao KK (2016) Divisive analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets. Alex Eng J 55(1):407–418
Polak RGE (1969) Note sur la convergence de méthodes de directions conjuguées. ESAIM 3(R1):35–43
Prakash J, Singh PK (2019) Gravitational search algorithm and k-means for simultaneous feature selection and data clustering: a multi-objective approach. Soft Comput 23(6):2083–2100
Prakash J, Singh PK (2015) Particle swarm optimization with k-means for simultaneous feature selection and data clustering. In: Second international conference on soft computing and machine intelligence (ISCMI2015), pp 74–78
Qian M, Zhai C (2013) Robust unsupervised feature selection. In: Proceedings of the twenty-third international joint conference on artificial intelligence, IJCAI'13, pp 1621–1627
Raftery AE, Dean N (2006) Variable selection for model-based clustering. J Am Stat Assoc 101(473):168–178
Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
Rui X, Wunsch ID (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678
Saha S, Acharya S, Kavya K, Miriyala S (2018) Simultaneous clustering and feature weighting using multiobjective optimization for identifying functionally similar mirnas. IEEE J Biomed Health Inform 22(5):1684–1690
Saha S, Ekbal A, Alok A, Spandana R (2014) Feature selection and semi-supervised clustering using multiobjective optimization. SpringerPlus 3:465
Saha S, Spandana R, Ekbal A, Bandyopadhyay S (2015) Simultaneous feature selection and symmetry based clustering using multiobjective framework. Appl Soft Comput 29:479–486
Samet R, Hancer E (2012) A new approach to the reconstruction of contour lines extracted from topographic maps. J Vis Commun Image Represent 23(4):642–647
Sarvari H, Khairdoost N, Fetanat A (2010) Harmony search algorithm for simultaneous clustering and feature selection. In: International conference of soft computing and pattern recognition, pp 202–207
Sheng W, Swift S, Zhang L, Liu X (2005) A weighted sum validity function for clustering with a hybrid niching genetic algorithm. IEEE Trans Syst Man Cybern B Cybern 35(6):1156–1167
Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF (2016) A new hybrid filter-wrapper feature selection method for clustering based on ranking. Neurocomputing 214:866–880
Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF (2019) A review of unsupervised feature selection methods. Artif Intell Rev. https://doi.org/10.1007/s10462-019-09682-y
Song M, Chen D (2018) An improved knowledge-informed NSGA-II for multi-objective land allocation (MOLA). Geo Spat Inf Sci 21(4):273–287
Storn R, Price K (1997) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 11(4):341–359
Swetha KP, Susheela Devi V (2012) Simultaneous feature selection and clustering using particle swarm optimization. In: Proceedings of the 19th international conference on neural information processing—volume part I, ICONIP'12. Springer, Berlin, pp 509–515
Tadesse MG, Sha N, Vannucci M (2005) Bayesian variable selection in clustering high-dimensional data. J Am Stat Assoc 100(470):602–617
Tsai CY, Chiu CC (2008) Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm. Comput Stat Data Anal 52(10):4658–4672
Turi R (2001) Clustering-based colour image segmentation. Ph.D thesis, Monash University, Australia
Vaithyanathan S, Dom B (1999) Generalized model selection for unsupervised learning in high dimensions. In: Proceedings of the 12th international conference on neural information processing systems, NIPS'99. MIT Press, Cambridge, MA, USA, pp 970–976

Vandenbroucke N, Macaire L, Postaire JG (2000) Unsupervised color texture feature extraction and selection for soccer image segmentation. vol 2

Wang H, Jing X, Niu B (2017) A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data. Knowl Based Syst 126:8–19

Wang H, Yan S, Xu D, Tang X, Huang T (2007) Trace ratio vs. ratio trace for dimensionality reduction. In: IEEE conference on computer vision and pattern recognition, pp 1–8

Wang L, Shen H (2016) Improved data streams classification with fast unsupervised feature selection. In: 17th international conference on parallel and distributed computing, applications and technologies (PDCAT), pp 221–226

Xie XL, Beni G (1991) A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 13(8):841–847

Xue B (2014) Particle swarm optimisation for feature selection. PhD thesis, Victoria University of Wellington, Wellington, New Zealand

Yang Y, Shen HT, Ma Z, Huang Z, Zhou X (2011) L2,1-norm regularized discriminative feature selection for unsupervised learning. In: Proceedings of the twenty-second international joint conference on artificial intelligence, vol 3, IJCAI'11. AAAI Press, pp 1589–1594

Ye J (2007) Least squares linear discriminant analysis. In: Proceedings of the 24th international conference on machine learning, ICML'07. ACM, New York, NY, USA, pp 1087–1093

Yun L, Bao-Liang L, Zhong-Fu W (2006) A hybrid method of unsupervised feature selection based on ranking. In: 18th international conference on pattern recognition (ICPR'06), vol 2, pp 687–690

Zhang T, Ramakrishnan R, Livny M (1997) Birch: a new data clustering algorithm and its applications. Data Min Knowl Discov 1(2):141–182

Zhang S, Wong H, Shen Y, Xie D (2012) A new unsupervised feature ranking method for gene expression data based on consensus affinity. IEEE/ACM Trans Comput Biol Bioinf 9(4):1257–1263

Zhao X, Xu G, Liu D, Zuo X (2017) Second-order de algorithm. CAAI Trans Intell Technol 2(2):80–92

Zhao Z, Liu H (2007) Spectral feature selection for supervised and unsupervised learning. In: Proceedings of the 24th international conference on machine learning, ICML'07. ACM, New York, NY, USA, pp 1151–1157

Zhao Z, Wang L, Liu H (2010) Efficient spectral feature selection with minimum redundancy. In: Proceedings of the twenty-fourth AAAI conference on artificial intelligence, AAAI'10, pp 673–678

Zhu QH, Yang YB (2018) Discriminative embedded unsupervised feature selection. Pattern Recognit Lett 112:219–225