Artificial Intelligence Review (2022) 55:2821–2846
https://doi.org/10.1007/s10462-021-10072-6

A survey on feature selection methods for mixed data

Saúl Solorio-Fernández[1] · **J. Ariel Carrasco-Ochoa**[1] · **José Francisco Martínez-Trinidad**[1]

Accepted: 11 September 2021 / Published online: 29 September 2021

Abstract
Feature Selection for mixed data is an active research area with many applications in practical problems where numerical and non-numerical features describe the objects of study. This paper provides the first comprehensive and structured revision of the existing supervised and unsupervised feature selection methods for mixed data reported in the literature. Additionally, we present an analysis of the main characteristics, advantages, and disadvantages of the feature selection methods reviewed in this survey and discuss some important open challenges and potential future research opportunities in this field.

Keywords Feature selection · Mixed data · Feature selection for mixed data · Dimensionality reduction

1 Introduction

Feature Selection, also known as *Attribute Selection* (Liu and Motoda 1998, 2007; Guyon et al. 2003), plays an essential role in Pattern Recognition, Machine Learning, Data Mining, Statistical Analysis, and many other research fields. *Feature Selection* is defined as the process that aims to reduce the set of features trying to approach, maintain, or even improve the performance of classification models regarding the performance obtained by using the whole set of features. Moreover, *Feature Selection* not only reduces the dimensionality of the data, but also it helps to remove non-relevant and redundant features, reducing the storage and processing requirements, avoiding the curse of dimensionality (Zhao and Liu 2011), and generating compact models with possibly better generalization capability (Pal and Mitra 2004).

✉ Saúl Solorio-Fernández
sausolofer@inaoep.mx

J. Ariel Carrasco-Ochoa
ariel@inaoep.mx

José Francisco Martínez-Trinidad
fmartine@inaoep.mx

[1] Computer Sciences Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro # 1, Tonantzintla, 72840 Puebla, Mexico

Over the last decades, many feature selection methods have been developed and applied to a wide variety of real-world problems (Jović et al. 2015; Li et al. 2017). Moreover, most of these methods have been introduced exclusively for numerical or non-numerical datasets, i.e., datasets where exclusively numerical or non-numerical features describe the objects of study. However, with the diversification of data types, mixed data[1], where both numerical and non-numerical features describe the objects, appears in many practical applications. For example, in clinical-medical decision-making systems, the coexistence of numerical and non-numerical features in the data is common. High dimensional data such as gene expression DNA microarray (Focant et al. 2011; van 't Veer et al. 2002), and medical datasets such as Heart (statlog) disease, Arrhythmia, Hepatitis, Horse colic, and Dermatology from the UCI Machine Learning Repository (Lichman 2013), are typically considered along with a few clinical features. These features can be numerical (e.g., blood pressure, temperature, blood glucose) or non-numerical (e.g., sex, smoker vs. non-smoker, blood type).

Another domain where mixed data are prevalent is in socio-economic and financial applications. In market research, the analysis of customer datasets that contain categorical features (e.g., type of a customer, preference, income group) and numerical features (e.g., age and the number of transactions, financial ratios) provide managers useful insights about customer behavior. Some examples of well-known public mixed data are the US consumer finances survey dataset (Hennig and Liao 2013), the German credit, Australian credit approval, and Boston Housing; these three last ones available in the UCI Machine Learning Repository (Lichman 2013). Furthermore, mixed data also appears in many other applications such as petroleum recovery (Lam et al. 2015), teaching assistant evaluation (Liang et al. 2012), intrusion detection (Liu 2012), network forensic (Ren et al. 2016), forest cover type (Ben Haj Kacem et al. 2015), Web-Based Learning Systems (Niu et al. 2015), and so on. Notwithstanding the foregoing, one of the major problems in this kind of data is the impossibility of using the same mathematical tools employed in feature selection methods developed exclusively for numerical or non-numerical data. Therefore, feature selection methods capable of processing mixed data are especially needed.

Currently, in the literature, there are several survey articles and books (Liu and Motoda 2007; Zhao and Liu 2011; Rudnicki et al. 2013; Bolón-Canedo et al. 2015) on feature selection for the supervised case (Kotsiantis 2011; Vergara and Estévez 2014), the semi-supervised case (Sheikhpour et al. 2017), the unsupervised case (Alelyani et al. 2013; Miao and Niu 2016; Solorio-Fernández et al. 2020; Hancer et al. 2020), as well as on feature selection in general (Liu et al. 2005; Chandrashekar and Sahin 2014; Li et al. 2016; Cai et al. 2018). Furthermore, some other reviews concentrate on describing feature selection applied to particular domains (Saeys et al. 2007; Mugunthadevi et al. 2011; Lazar et al. 2012; Bharti and kumar Singh 2014; Ang et al. 2016; Lee et al. 2017; Liang et al. 2018; Deng et al. 2019; Remeseiro and Bolon-Canedo 2019; Zhang et al. 2019). However, none of them have focused specifically on feature selection for mixed data. This survey aims to fill this gap and provide a comprehensive and structured revision of the feature selection methods for mixed data reported in the literature. Moreover, this survey also highlights the general advantages and disadvantages of feature selection methods from each approach and also discusses some open challenges and potential research opportunities in this field of research. To the best of our knowledge, this is the first comprehensive survey of feature

[1] Also called heterogeneous or assorted data.

selection methods for mixed data. Specifically, the contributions of this paper can be summarized as follows:

- It presents the first survey that is focused on feature selection for mixed data which reviews the more relevant and recent supervised and unsupervised state-of-the-art feature selection methods.
- It presents a categorization of the feature selection methods for mixed data reviewed as well as a general discussion about their respective advantages and disadvantages.
- It provides a thorough analysis of the main characteristics of the reviewed feature selection methods for mixed data and highlights some open challenges and potential research directions.
- It serves as a guide to aid researchers, practitioners, and academics, who focus on developing Machine Learning algorithms and those that perform data analysis in classification and clustering tasks applied to mixed data.

The layout of this survey is as follows: Sect. 2, introduces the different types of feature selection, including the main paradigms, approaches, and the most common feature selection evaluation strategies. Sections 3 and 4 review the supervised and unsupervised feature selection methods for mixed data introduced in the literature. The analysis and discussion of the main approaches and methods reviewed in this survey are presented in Sect. 5. Finally, Sect. 6 provides our concluding remarks as well as some open challenges and future research directions.

2 Feature selection types and their evaluation

According to the availability of information in the data, feature selection methods can be classified as *supervised* (Kotsiantis 2011; Tang et al. 2014), *semi-supervised* (Sheikhpour et al. 2017) and *unsupervised* (Fowlkes et al. 1988; Dy and Brodley 2004; Alelyani et al. 2013). The former require a set of labeled[2] data (supervised dataset) to perform the feature selection, where the main objective is to find a subset of features that maximizes some function of prediction or accuracy. Semi-supervised methods, on the other hand, require only that some objects be labeled, and the general objective is to use the label information of labeled data and local structure of both labeled and unlabeled data to evaluate the features' relevance (Sheikhpour et al. 2017). Meanwhile, Unsupervised Feature Selection (UFS) methods do not require a labeled dataset, and the objective is to select a subset of informative or relevant features that allows finding good cluster structures in the data (Dy and Brodley 2004).

2.1 Approaches

Feature selection methods (supervised, semi-supervised, and unsupervised) can be categorized into three main approaches as *filter*, *wrapper*, and *hybrid*, according to the evaluation criterion used for selecting features (Liu et al. 2005; Alelyani et al. 2013). Methods

[2] The label assigned to each object in the dataset can be a category, an ordered value, or a real value, depending on the specific task.

based on the filter approach select features using only the intrinsic properties of the data, applying evaluation measures such as variance, similarity among features, consistency, entropy, and others. In turn, filter methods can be subdivided into univariate and multivariate (Alelyani et al. 2013). Univariate methods (also known as ranking-based methods) evaluate every single feature to get an ordered list (ranking). Meanwhile, multivariate filter methods evaluate the relevance of the features with respect to other features. A typical filter method comprises two basic components; a feature search strategy and a feature evaluation criterion. In the feature search strategy, a feature subset is generated, and then this is evaluated through a predefined intrinsic quality measure. This process is repeated until some pre-established stop criterion is met.

On the other hand, methods based on the wrapper approach evaluate feature subsets depending on their performance under a specific classification model (classification/clustering algorithm). Methods based on the wrapper approach can be subdivided as sequential, bio-inspired, and iterative, according to the feature search strategy used (Solorio-Fernández et al. 2020). In the former, features are added or removed sequentially, and a supervised or unsupervised classification model is built and evaluated through some quality criterion using the selected features. Meanwhile, bio-inspired methods try to incorporate randomness into the search process, aiming to escape from local optima. Finally, iterative methods address the feature selection procedure by casting it as an estimation problem and thus avoiding a combinatorial search. A general feature selection method based on the wrapper approach consists of three basic components: a search strategy, a classification model, and a feature evaluation criterion. In the first component, a candidate feature subset is generated based on a given search strategy; then, in the second component, a classification/clustering algorithm is applied to the data described through the candidate feature subset. In the final component, the results are evaluated according to a feature evaluation criterion. The subset that best fits the evaluation criterion will be chosen from all the evaluated candidates.

Finally, methods based on the hybrid approach can be either formed by combining two different approaches (e.g., filter and wrapper) or a combination of two methods under the same approach (Ang et al. 2016). The most common hybrid methods are those combining methods from the filter and wrapper approaches. A typical hybrid method goes through the following steps: (1) applying a filter criterion to select different candidate subsets or to produce a feature ranking. Then, (2) evaluating the quality using a classification model for each candidate subset or some feature subsets built from the feature ranking. In the last step, (3) the subset with the highest quality is selected.

2.2 Performance evaluation

In the literature, there are two ways commonly used to assess the performance of feature selection methods (Li et al. 2016): (1) in terms of clustering results through external evaluation measures; and (2) in terms of supervised classification results of a particular supervised classifier. The former way is used exclusively to evaluate unsupervised feature selection (UFS) methods; meanwhile, the second way can be used for the supervised, semi-supervised, and unsupervised cases.

In order to evaluate the performance of UFS methods in terms of clustering results, first, a UFS method is applied over the original dataset to select/rank features. Then, a clustering algorithm is applied over the data described just by the selected features (reduced data), and the clustering results are assessed through an external evaluation measure. For

assessing clustering results, the *Clustering Accuracy* (ACC) and the *Normalized Mutual Information* (NMI) (Xu et al. 2003; He et al. 2005) are commonly used. On the other hand, for assessing the performance of feature selection methods regarding supervised classification results, the *classification accuracy* of a particular classifier is commonly applied as follows: first, the whole dataset is usually divided into two parts—a training set and a test set. Then, a feature selection method[3] is applied on the training set to obtain a feature ranking or a feature subset. Later, after training the classifier using the training set on the selected features, the test set on the selected features is used for assessing the classifier through its accuracy or error rate. To get more reliable results, a k-fold cross-validation technique is usually applied, and the final classification performance is reported as an average over the k folds. The higher the average classification accuracy, the better the feature selection method.

Other ways for evaluating feature selection methods that have been used in the literature include evaluation in terms of *correctness of the selected features*, *feature redundancy*, *degree of dimensionality reduction*, and feature selection *runtime*. Evaluation of the correctness of the selected features consists in quantifying the number of relevant features selected by a feature selection method through an information retrieval measure such as precision, recall, or F-measure. This evaluation is done using synthetic or real-world datasets where the actual relevant features are known a priori. On the other hand, the evaluation in terms of feature redundancy is used by methods that consider eliminating redundancy. The *redundancy rate* (Zheng et al. 2010) and *Representation Entropy* (Devijver and Kittler 1982; Mitra et al. 2002; Rao and Sastry 2012) are the most used feature redundancy measures. Evaluations in terms of degree of dimensionality reduction and runtime consider only the size of the selected subset and the runtime spent for the selection.

3 Supervised feature selection for mixed data

Feature Selection for mixed data in the supervised context has been applied in practical problems for many years (Ruiz-Shulcloper and Abidi 2002; Ruiz-Shulcloper 2008). Classical and relevant filter feature selection methods such as CFS (Hall 2000), FCBF (Yu and Liu 2003), ReliefF (Kononenko 1994), and mRMR (Peng et al. 2005) can handle mixed data. Correlation Feature Selection (CFS) (Hall 2000) is a multivariate method that quantifies the features' relevancy by evaluating the *merit* of feature subsets by using a correlation-based evaluation measure and handling both numerical and non-numerical features as non-numerical (numerical features are discretized). Likewise, Fast Correlation-Based Filter (FCBF) (Yu and Liu 2003) selects relevant and non-redundant features applying a pairwise-correlation strategy using an entropy-based measure, where numerical features are binned through the discretization method introduced in Fayyad and Irani (1993). ReliefF (Kononenko 1994) for its part, deals with this problem using the Hamming distance for non-numerical features and the Euclidean distance for numerical ones. On the other hand, minimal-Redundancy-Maximal-Relevance (mRMR) (Peng et al. 2005) handles mixed data estimating the mutual information of non-numerical features directly from the data. Meanwhile, for numerical features, the authors suggest using a discretization method as a preprocessing

[3] For the case of UFS methods, class labels are not used in this step.

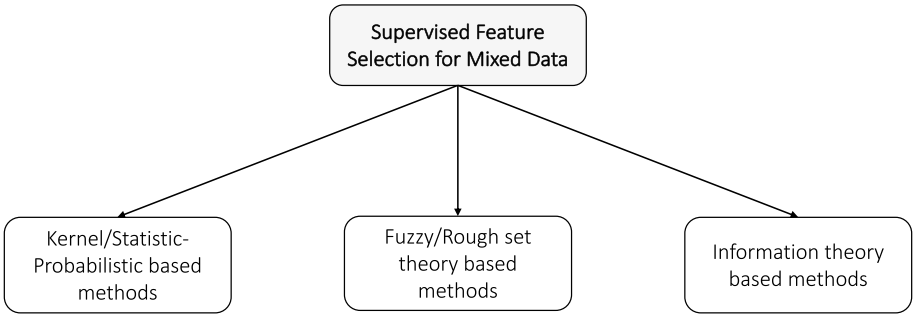

Fig. 1 Categorization of supervised feature selection methods specially developed for mixed data

step or a density estimation method like Parzen windows for the mutual information approximation.

Other more recent supervised feature selection methods such as UFT (Wei et al. 2015), mDSM (Sharmin et al. 2019), and the method introduced in Doan et al. (2020) perform feature selection in mixed data transforming features to a particular type. In Wei et al. (2015), a novel Unsupervised Feature Transformation (UFT) method, which can transform non-numerical features into numerical ones, was developed and tested jointly with some supervised feature selection methods capable of handling mixed data. The idea of UFT is to find a numerical representation to substitute the original non-numerical feature values using mutual information-based concepts. After, once the feature transformation has been done, conventional selectors such as mRMR (Peng et al. 2005) and PWFS (Kwak 2002) are employed to select the final feature subset. On the other hand, in Sharmin et al. (2019), a filter method called modified Discretization and feature Selection based on Mutual information (mDSM) was proposed. This method simultaneously applies feature selection and discretization, dividing this procedure into two major parts. In the first part, the idea is to find the discretization level of each feature using a Joint Bias corrected Mutual Information measure (proposed in this same work), obtaining a feature ranking according to the mutual information value associated with each feature. In the second part, following the feature ranking, mDSM evaluates candidate feature subsets using a modified unbiased Mutual Information criterion for selecting the final feature subset. Finally, in Doan et al. (2020), a wrapper method for selecting the best possible optimal features by evaluating the performance of a chosen classification algorithm was proposed. This method requires determining both a classification algorithm and a performance metric to build a learning model by adding or removing features iteratively. It is important to mention that, for mixed data, categorical features are coded using a one-hot encoding procedure to convert categorical values to n-dimensional one-hot encoded features.

On the other hand, several supervised feature selection methods specially developed for mixed data have also been introduced in the literature. Figure 1 shows a taxonomy of supervised feature selection methods specially developed for mixed data reviewed in this survey. We have classified them into three main categories: Kernel/Statistic-Probabilistic, Fuzzy/Rough set theory, and Information theory, based methods. In the following, we will briefly describe and discuss the methods in each considered categorization.

3.1 Kernel/statistic-probabilistic based methods

The basic idea of Kernel/Statistic-Probabilistic methods is to evaluate the relevancy of features using kernel functions, probability/statistical measures, or different correlation measures (one for each type of feature) to quantify the degree of relationship or association among features. This kind of methods although were among the first to be applied to mixed data for feature selection in the supervised context, they are still being developed today with many applications in real-world problems. Examples of this kind of method are MFS (Tang and Mao 2005, 2007), RFE (Paul and Dupont 2014; Paul et al. 2015), SSFSM (Solorio-Fernández et al. 2019), ECMBF (Jiang and Wang 2016), and Mixed-MB (Lee et al. 2020).

Mixed Forward Selection search algorithm (MFS) (Tang and Mao 2005, 2007) is an early hybrid method for mixed data where features are evaluated through the *join error probability* (Devijver and Kittler 1982); which is defined as the weighted sum of conditional error probabilities of continuous features given nominal features. MFS performs feature selection as follows: first, features are divided into two groups *Num* and *Nom*, denoting the numerical and non-numerical features, respectively. Then, each group of data is ranked in parallel (using the join probability error as the quality measure) through a sequential forward selection procedure in conjunction with an in-depth search, generating in this way several step-optimum candidate feature subsets at each step. Afterward, to select the best feature subset, the results of a classifier trained using each subset of features are employed. The stopping criterion in MFS is either a predefined feature subset size or a cross-validation error rate.

On the other hand, Recursive Feature Elimination (RFE) (Paul and Dupont 2014; Paul et al. 2015) is a wrapper kernel-based method that performs feature selection using three components: (1) a dedicated kernel that can handle mixed data, (2) a feature search strategy, and (3) a classifier (usually SVM) which allows quantifying the importance of each feature through an objective function that uses the dedicated kernel. RFE iteratively builds a feature ranking by removing the least relevant feature at each step. The feature relevancy is quantified by the influence on the margin when a particular feature is removed, in such a way that those features that do not decrease much the margin size are considered as not relevant, and they are eliminated; hence, the feature importance in RFE is thus evaluated with respect to the separating hyperplane. Another kernel-based method but developed under the filter approach is Supervised Spectral Feature Selection Method for mixed data (SSFSM) (Solorio-Fernández et al. 2019). SSFSM is a univariate method based on Spectral Feature Selection and a new supervised kernel to select relevant features in mixed datasets. The SSFSM's idea is to apply a leave-one-out feature elimination strategy by using the spectral gap score (Solorio-Fernández et al. 2017) to quantify the consistency of each feature. This work introduces a new supervised kernel that uses both the information contained in the features and the information provided by the target objective. In SSFSM, features are ranked from the most to the least relevant using the corresponding spectral gap scores.

Efficient Correlation Measure Based Filter (ECMBF) (Jiang and Wang 2016) is a statistic filter feature selection method based on correlation analysis whose objective is to remove irrelevant and redundant features. In ECMBF, feature selection is performed in two steps: first, a set of relevant features is selected by measuring the correlation between each feature and the class. Then, a final feature subset with low redundancy is selected by measuring the correlation among features. To quantify the relationship

between numerical and non-numerical features, a new correlation measure based on the class separation theory was introduced; meanwhile, linear correlation and symmetrical uncertainty were used for numerical and non-numerical features, respectively. Finally, another statistical-based method developed under the filter approach but can be applied to mixed data in both classification and regression problems is Mixed-MB (Lee et al. 2020). The basic idea is to apply a series of conditional independence tests (CI tests) into a Markov Blanket (MB) search algorithm (Koller and Sahami 1996) in order to find a minimal set of relevant and non-redundant features. Mixed-MB introduces a generalized CI test based on the likelihood-ratio (LR) (Wilks 1938), which can handle mixed data without information loss, and it can be embedded in the INTER-IAMB algorithm (Tsamardinos et al. 2003) for feature selection. In this method, the LR test is used to determine the CI, and, in each model, categorical features are coded as dummy variables.

3.2 Fuzzy/rough set theory based methods

Fuzzy/Rough set theory-based methods evaluate features using fuzzy relations or equivalence classes (also called granules). The basic idea is to determine the minimal subset of the selected features containing the maximal number of elemental granules. Examples of these methods are FarVPN (Hu et al. 2008a, b), EFSH (Wang and Liang 2016), FSMSD (Kim and Jun 2018), MIFSA/MIFSD (Sang et al. 2020) and the feature selection methods introduced by Chen and Yang (2014) and Zhang et al. (2016).

Forward attribute reduction based on Variable Precision Neighborhood model (FarVPN) (Hu et al. 2008a, b) is a filter method based on granulation and approximation. The main idea is extending Pawlak's rough set model (Pawlak 1982) to mixed data, where numerical features induce a set of δ-neighborhood or k-nearest-neighbor granules; meanwhile, non-numerical features generate crisp equivalence relations and equivalence classes on the sample spaces. Then, the granules are used to approximate the decision class in the framework of rough sets and select a final feature subset. A forward greedy search algorithm was applied to find a minimal feature subset (a *reduct*), which can keep classification ability. A method related to FarVPN is Efficient Feature Selection algorithm for large-scale Hybrid decision tables (EFSH) (Wang and Liang 2016), which proposes to divide the data into n subsets (subtables) using a new sampling algorithm based on the idea of decomposition and fusion. Then, from the subtables generated in the previous step, the method gets a final feature subset using the same algorithm introduced in FarVPN (Hu et al. 2008a) for computing *reducts*.

Another two filter methods based on a rough set model are FSMSD (Kim and Jun 2018) and MIFSA/MIFSD (Sang et al. 2020). Feature Selection for Mixed-type data with feature Space Decomposition (FSMSD) performs feature selection in three steps: (1) ranking features separately by sorting the numerical and categorical features into descending order. The numerical features are ranked using the ERGS (Chandra and Gupta 2011) algorithm; meanwhile, categorical features are ranked using an equivalence relation. (2) Using the join probability error (Tang and Mao 2005, 2007) as a quality measure for selecting a candidate feature (numerical or categorical) and decide if such feature is placed in the final feature subset. Finally, (3) repeating step number 2 until no more features are left to evaluate. On the other hand, in Sang et al. (2020) two feature selection methods called Matrix-based Incremental Feature Selection algorithm (MIFSA) and Matrix-based Incremental Feature Selection algorithm while Deleting multiple objects (MIFSD) were proposed. In this work,

the authors aim to apply feature selection for heterogeneous dynamic ordered data, i.e., mixed datasets that change over time. The general idea is inspired by the incremental filter dominance-based neighborhood rough set approach (Greco et al. 2001), where a new dominance conditional entropy measure and a feature selection strategy based on this measure were proposed.

In Chen and Yang (2014), it was introduced a filter method which performs feature selection in mixed data by composing classical rough set theory (Pawlak 1982) and fuzzy rough set models (Jensen and Shen 2004). The idea is to characterize the *discernibility* of the features regarding the decision labels based on the mutual effects between numerical and non-numerical features, i.e., the ability to discern if a numerical feature can be substituted by a non-numerical one. This substitution comes down to finding the "ϵ-reducts" from the discernibility matrix derived from the data. To find ϵ-reducts, a fast search algorithm based on the discernibility matrix (Chen et al. 2012) was used. Finally, another feature selection method, also based on Fuzzy/Rough set theory but developed under the hybrid approach, was proposed by Zhang et al. (2016). In this method, the main objective is to use a λ-conditional entropy measure (filter stage) for generating a feature ranking into a general fuzzy decision system described by numerical and non-numerical features. Then, in the wrapper stage, using a heuristic feature selection algorithm combined with a classifier, the classification accuracy is used for selecting the best "ϵ-approximate" *reduct* (the best feature subset).

3.3 Information theory based methods

As their name implies, Information theory-based methods evaluate features using measures taken from the Information theory (Cover and Thomas 2006) such as mutual information, conditional mutual information, or entropy. Examples of Information theory-based methods are Hybrid-MI (Doquire and Verleysen 2011a), Mixed-MI (Doquire and Verleysen 2011b), HFS (Liu et al. 2013), RnR-SSFSM (Solorio-Fernández et al. 2020), and the feature selection method introduced by Coelho et al. (2016).

Hybrid-MI (Doquire and Verleysen 2011a) is a hybrid method that independently ranks numerical and non-numerical features before recombining them according to a classifier's accuracy. More precisely, the features of each type are first ranked independently, using the Mutual Information (MI) criterion (Peng et al. 2005) producing, two independent lists; after, the lists are combined according to Naive Bayes and kNN classifiers' accuracy. Hybrid-MI estimates the MI from the data using a family of estimators based on the nearest neighbors paradigm (Kraskov et al. 2004). A further version of Hybrid-MI called Mixed-MI was proposed by the same authors in Doquire and Verleysen (2011b); the idea was to extend the Hybrid-MI method to regression problems by introducing some modifications to the MI estimator. Another information-based method developed under the hybrid approach is HFS (Liu et al. 2013). Hybrid Feature Selection (HFS) introduces a new normalized correlation measure based on a Parsen-window estimator for computing the Mutual Information between numerical and non-numerical features; then the mRMR criterion (Peng et al. 2005) is applied for filtering the first p features. At the last step, the final feature subset is obtained by minimizing the estimation accuracy of a Case-Based Reasoning (CBR) wrapper model.

On the other hand, Relevant and non-Redundant Supervised Spectral Feature Selection method for Mixed-data (RnR-SSFSM) (Solorio-Fernández et al. 2020) is a filter method that combines Spectral Feature Selection (Zhao and Liu 2011) and Information-theory

based redundancy analysis (Yu and Liu 2004). RnR-SSFSM performs feature selection in two stages; first, features are ranked and selected using the feature selection method for mixed data introduced in Solorio-Fernández et al. (2017). Then, RnR-SSFSM selects a feature subset of relevant and non-redundant features through pairwise correlation analysis using an entropy-based measure. It is worth mentioning that for using this entropy-based measure, numerical features are discretized using the supervised MDL discretizer introduced in Fayyad and Irani (1993).

Finally, in Coelho et al. (2016), a filter feature selection strategy along with a Mixed Mutual Information estimator (MMI) based on the Kraskov method (Kraskov et al. 2004) that can deal with continuous and discrete features was proposed. The goal is to apply forward and backward search strategies to get the feature subsets with the largest MMI regarding the objective concept (class labels). In this method, since the evaluated feature subsets have non-comparable dimensions, the authors use a permutation test (François et al. 2006) as a stopping criterion.

3.4 Summary

In this section, we reviewed feature selection methods capable of handling mixed data in the supervised context. First, we revised classical methods of the state-of-the-art that, although they were not designed exclusively for mixed data, they can process this type of data. These methods (especially the filter multivariate ones) usually obtain excellent results in practical problems since they evaluate features jointly rather than individually. However, the main drawback is that they apply feature transformation procedures or used different evaluation measures, which bring about important inconveniences, as we will discuss later in Sect. 5.2.

On the other hand, we also reviewed those methods specially designed to process mixed data, which we divided into three major categories: Kernel/Statistic-Probabilistic, Fuzzy/Rough set theory, and Information theory-based methods. The Kernel/Statistic-Probabilistic and the Fuzzy/Rough set theory-based methods were among the first to address supervised feature selection on mixed data. Kernel/Statistic-Probabilistic based methods are straightforward and can be applied to high dimensional datasets (especially those filter univariate). Moreover, these methods provide results with good stability (Paul et al. 2015); however, it is not clear how to determine the best kernel function or a suitable statistical/probabilistic correlation measure for the different types of data. Fuzzy/Rough set theory-based methods, for their part, provide a natural framework for dealing with mixed data because, for different kinds of features, fuzzy rough sets can be defined to measure the similarity between objects (Zhang et al. 2016). However, they are time-consuming due to *reducts* computations, limiting them to apply only to low dimensional datasets with few objects. Moreover, some of these methods (Hu et al. 2008a, b; Wang and Liang 2016) use the concept of granulation to handle numerical data. Nevertheless, according to Liu et al. (2013), the granulation's shortcoming is that the scale parameter is not easy to determine, which leads to instability. Finally, Information theory-based methods are popular due to their capability to capture non-linear relationships between features (Vergara and Estévez 2014; Su and Liu 2018). Another attractive property of these methods is that they can be applied in both categorical and numerical features. However, a remarkable drawback for most of these methods (Doquire and Verleysen 2011b; Liu et al. 2013; Coelho et al. 2016) is that for computing mutual information (MI) among continuous numerical features in

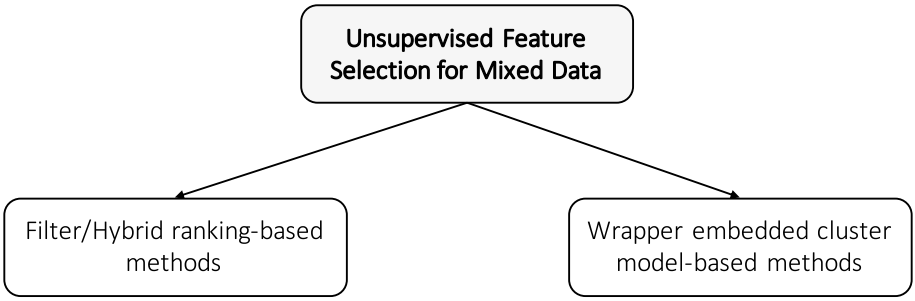

Fig. 2 Categorization of unsupervised feature selection methods for mixed data

mixed data, a reliable estimate of MI is challenging to obtain whenever fewer samples than dimensions are available (Paul and Dupont 2014).

4 Unsupervised feature selection for mixed data

In the literature, we can found several classical and recent Unsupervised Feature Selection (UFS) methods based on different approaches and frameworks (Alelyani et al. 2013; Li et al. 2016; Solorio-Fernández et al. 2020; Hancer et al. 2020). Notwithstanding the foregoing, feature selection for mixed data in unsupervised contexts has been much less studied than its supervised counterpart. Instead, this problem in practice has been addressed using feature transformation techniques such as encoding, discretization, or directly embedding feature selection in wrapper cluster based-models (Fop and Murphy 2018). Moreover, to the best of our knowledge, few UFS methods are capable of processing mixed data. This section presents a revision of these methods, which we have divided into Filter/Hybrid ranking-based methods and Wrapper Embedded cluster model-based methods. Figure 2 shows the taxonomy of the UFS methods reviewed in this survey.

4.1 Filter/hybrid ranking-based UFS methods for mixed data

One of the first works developed under the filter approach, where the authors mention the proposed UFS method allows processing mixed data, was introduced by Dash et al. (1997). In this work, the authors proposed a filter solution called Sequential backward selection method for Unsupervised Data (SUD), which is based on a measure of "entropy of similarities". The general idea consists of measuring the entropy of the data based on the fact that if most of the distances between pairs of objects are very close or very far, the entropy is low; and it is high if most of the distances between pairs of objects are close to the average distance. Therefore, if the data has low entropy, there are well-defined cluster structures; on the contrary, the cluster structures are not well-defined if the data has high entropy. In SUD, each feature's relevance is quantified by the entropy measure applying a leave-one-out sequential backward strategy, i.e., each feature, in the whole set of features, is removed in turn, and the entropy of the dataset without that particular feature is computed. The final result is a feature ranking ordered from the most relevant feature (which produces the highest entropy when it is removed) to the least relevant one. It is noteworthy that in this method for handling mixed data, the authors recommend performing a feature

transformation by applying the Chi2 (Liu and Setiono 1995) discretization method on the numerical features before using the Hamming distance.

A later USF method, related to the previous one, also capable of handling mixed data but under the hybrid approach, was introduced by Dash and Liu (2000). According to the authors, Dash and Liu's method can be applied to mixed data by discretizing the numerical features in the filter stage and applying an encoding method in the wrapper stage. Dash and Liu's method is based on the entropy measure proposed in Dash et al. (1997) (filter stage) jointly with the internal scatter separability criterion (Dy and Brodley 2004) (wrapper stage). First, in the filter stage, the authors apply a leave-one-out search strategy using the entropy measure to obtain a list of features sorted according to the degree of disorder that each feature generates when it is removed from the whole set of features. Then, in the wrapper stage, following the feature ranking produced in the filter stage, a forward selection search is applied jointly with the *k*-means algorithm to build clusters, which are evaluated by the scatter separability criterion. As a result, Dash and Liu's method provides the subset of features that reaches the highest value for the scatter separability criterion in the forward selection search.

Two more recent ranking-based UFS methods are USFSM (Solorio-Fernández et al. 2017) and the method proposed by Chaudhuri et al. (2021). Unsupervised Spectral Feature Selection method for Mixed-data (USFSM) is a filter univariate Spectral Feature Selection method that weights the features according to their capability for defining good cluster structures in the data. The feature evaluation is performed by using a similarity function (kernel function) for mixed data, a spectral-based feature evaluation function, and a leave-one-out feature elimination strategy that, all together, allow to rank features in the data independently of their type. The idea is to weight features according to their consistency (Zhao and Liu 2007) by analyzing the changes in the spectrum distribution (spectral gaps) of the Normalized Laplacian matrix (Luxburg 2007) when each feature is excluded from the whole set of features separately. In USFSM, features are sorted in descending order according to their respective weights, i.e., according to their spectral gap score values. On the other hand, the method introduced by Chaudhuri et al. (2021) is a hybrid multivariate selector that utilizes entropy, mutual information, and a normalized clustering assessment measurement to select informative and non-redundant features on mixed datasets. This method selects a feature subset in two phases. First, highly redundant features are removed using entropy to identify individual features' information, and mutual information to find feature-feature correlations. Lastly, in the second phase, the *k*-prototypes (Huang 1997, 1998) clustering algorithm and a normalized Calinski-Harabasz evaluation index are employed to obtaining a relevant reduced ranked feature subset. It is important to mention that numerical features are treated as categorical integers for computing entropy and mutual information.

4.2 Wrapper embedded cluster model-based UFS methods for mixed data

Wrapper methods based on embedded models have also been frequently proposed in the unsupervised context to handle mixed data in the last years. A wrapper UFS method based on an evolutionary algorithm that simultaneously performs clustering and feature selection was introduced by Dutta et al. (2014). In this method, feature selection is performed while the data are clustered using a Multi-Objective Genetic Algorithm (MOGA). MOGA proposes a multi-objective fitness function that minimizes the intra-cluster distance (Homogeneity) and maximizes the inter-cluster distance (Separation). In MOGA, each chromosome

represents a solution composed of a set of c cluster centroids described by a subset of features. For continuous features, the centroid is computed as the mean in the cluster, while for categorical features, the centroid is computed as the mode (the most frequent feature value). First, the set of features used for each centroid in each chromosome is randomly selected in the initial population. Then, for reassigning cluster centroids, MOGA uses the k-prototypes clustering algorithm (Huang 1997, 1998), which obtains its inputs from the initial population generated in the previous step. Afterward, for every chromosome in the population, MOGA calculates the fitness function's value, and those chromosomes lying on the Pareto optimal front (dominant chromosomes) are selected for the next generation. Later, the crossover, mutation, and substitution operators are applied to the selected population, and the process is repeated until a pre-specified stop criterion is met (maximum number of generations). Once the stop criterion has been met in the final stage, MOGA returns the set of chromosomes in the Pareto population as the best solutions, jointly with the clusters produced by each chromosome.

On the other hand, a wrapper iterative-embedded UFS method for mixed data called CRAFT (ClusteR-specific Assorted Feature selecTion) was introduced by Garg et al. (2016). This method defines a generative hierarchical Bayesian model for clustering with cluster-specific feature selection, i.e., this method can select a different set of features for each cluster depending on an a priori defined parameter[4] that can be adjusted for the desired balance between global and local feature selection. CRAFT's basic idea relies on assuming that the objects in a cluster should closely agree on the features selected for that cluster; this turns out in an objective function based on the cluster's entropy for non-numerical features and variance for numerical ones. CRAFT performs the following steps repeated in each iteration until the cluster assignments do not change. (1) Compute the distance of each object to the cluster centers, (2) choose which cluster each object should be assigned to (and create new clusters if needed), and (3) recompute (update) the centers. The update of the cluster centers and the feature selection are performed at the end of each iteration. CRAFT output is a feature subset, the number of clusters, and the clusters assignment produced using the selected features. Finally, it is worthy to note that the authors mention that they do not know whether CRAFT is guaranteed to converge to a solution.

Two more recent iterative-embedding methods that have been proposed for feature selection in mixed data under the wrapper approach are VarSelLCM (Marbac and Sedki 2017, 2019) and DPM-MCMC (Storlie et al. 2018), both based on finite mixed models. VarSelLCM is a constrained-based model that assumes that features are independent within components (clusters); according to the authors, this assumption facilitates model selection since it allows easily to implement a modified version of the Expectation-Maximization (EM) algorithm (Dempster et al. 1977). VarSelLCM considers two types of features: the set of the relevant features (having different distribution among components) and the set of the irrelevant features (having the same distribution among components), which are considered independent of the relevant ones. A binary vector $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)$ defines the role of each feature, where $\omega_j = 0$ if the feature j is irrelevant and $\omega_j = 1$ otherwise. In order to get the relevant features as well as the parameters of the mixed model, an information criterion like BIC (Schwarz 1978), AIC (Akaike 1998) or MICL (Marbac and Sedki 2017, 2019) is optimized by using a modified version of the EM algorithm (Green 1990). Lastly, once the model has converged, VarselLCM returns the set of features composed of

[4] A parameter given by the user in the range (0, 1) that specifies the average fraction of features per cluster.

those features with $\omega = 1$ (relevant features), the clusters, and the optimized model parameters. DPM-MCMC (Dirichlet Process Model-Markov chain Monte Carlo), for its part, is a Bayesian nonparametric wrapper method that simultaneously estimates the number of clusters, cluster membership, and feature selection. The non-numerical features and the numerical ones are handled with a Gaussian latent variable approach. Unlike VarselLCM, DPM-MCMC performs feature selection via stochastic search variable selection (SSVS) (George and McCulloch 1993, 1997) instead of using the information criteria BIC, AIC or MICI. In DPM-MCMC, the model parameters are sampled by using a Markov chain Monte Carlo technique for inference on the Dirichlet Process Model. The informative (relevant) features are represented by a vector $\boldsymbol{\gamma}$ of binary values such that $\gamma = 1$ if the feature is informative, and $\gamma = 0$ otherwise. The parameters of the model, as well as the vector $\boldsymbol{\gamma}$, are updated iteratively in the Metropolis-Hastings (MH) step (George and McCulloch 1997) of the MCMC process until a pre-defined number of iterations is met. Afterward, once the model has converged, DPM-MCMC returns the clusters, the set of parameters of the final model, and the binary vector $\boldsymbol{\gamma}$ representing the selected features.

4.3 Summary

In this section, Unsupervised Feature Selection (UFS) methods for mixed data were reviewed and classified as Filter/Hybrid ranking-based and Wrapper Embedded cluster model-based methods. The former commonly try to take advantage of the ranking in the filter stage to select good candidate feature subsets to evaluate in the second stage. However, their main drawback is that most of these methods perform an a priori discretization or codification of features before being applied on mixed data, resulting in some problems, as we will see later in Sect. 5.2. On the other hand, Wrapper Embedded cluster model-based methods often get good results because the clustering algorithm used helps to evaluate the features and find informative feature subsets. Nevertheless, most of these methods optimize an internal validation index or an information criterion, which could bias the feature selection results, selecting in many cases, all or just one feature due to the so-called "Bias of Criterion Values to Dimension" (Dy and Brodley 2004). Moreover, most of these methods were not proposed specially for feature selection, and their main objective is clustering, and feature selection is a secondary task. Another important inconvenience is that these methods were not designed to be applied with any other clustering algorithm than the one used in the embedded model, reducing in this way their application scope.

In summary, from our revision, we can conclude that few studies address the problem of unsupervised feature selection specifically for mixed data. Moreover, most of the current UFS methods do not really solve the feature selection in mixed data because they have several characteristics that adversely affect the quality of their results. Therefore, new UFS methods for mixed data that solve the problems above discussed are still needed.

5 Analysis and discussion

In the previous sections, supervised and unsupervised feature selection methods for mixed data were categorized and reviewed according to their approach and type. This section analyzes and presents a general discussion about the main advantages and disadvantages of the different approaches and summarizes the main characteristics of feature selection methods for mixed data reviewed in this survey.

Table 1 General advantages and disadvantages of feature selection methods regarding their approach

Approach	Advantages	Disadvantages
Filter	Faster than wrapper	
	Scalable	
	Independent of the classification/clustering algorithm	Ignoring interaction with classification/clustering algorithms
	Parallelizable	
	Better generalizable property	
Wrapper	Interact with the classification/clustering algorithm to be used	Risk of overfitting
	Can model feature dependencies	High computational cost
	Higher performance accuracy than filter	The selection is specific for the used classification/clustering algorithm
Hybrid	Interact with the classification/clustering algorithm to be used	
	Less time consuming than wrappers	The selection is specific for the used classification/clustering algorithm
	Can model feature dependencies	
	Less prone to over-fitting than wrapper	
	Higher performance accuracy than filter	

5.1 Advantages and disadvantages of general approaches

In Table 1 we summarize the main advantages and disadvantages of feature selection methods categorized by approach. From this table, it is possible to observe that, in general, there is not a feature selection approach that is the best for all problems; every approach has its pros and cons. On the one hand, filter methods are straightforward and independent of any classification/clustering algorithm; therefore, they can provide more general solutions (Ang et al. 2016). Moreover, filter methods, specifically the univariate ones, can effectively identify irrelevant features and tend to be fast and scalable. However, they cannot remove redundant features because they do not consider possible dependencies between features. Meanwhile, multivariate filter methods can handle redundant and irrelevant features since they evaluate features jointly. However, multivariate methods have a higher computational cost compared to univariate ones.

On the other hand, wrapper methods usually obtain better performance results than the filter methods because they consider the feature dependencies and directly incorporate the bias of a particular classification/clustering algorithm. However, the main disadvantage is that they typically have a high computational cost because the defined classification/clustering algorithm must be re-executed each time. Furthermore, wrapper methods are more prone to over-fitting than the filter method because the classification/clustering algorithm is repeatedly called to evaluate each subset. Hybrid methods, for their part, commonly attempt to inherit the advantages of both approaches (filter and wrapper) by combining their complementary strengths. They use different evaluation criteria to improve the efficiency and prediction performance with better computational performance. As a result, hybrid methods usually produce better quality solutions than the methods of the filter approach, but they are less efficient. Compared to the wrapper approach, the hybrid methods are more efficient (Aggarwal and Reddy 2013), but they usually produce lower quality solutions.

It is worth mentioning that in the literature, several authors (Saeys et al. 2007; Jović et al. 2015; Chandrashekar and Sahin 2014; Li et al. 2016; Ang et al. 2016; Li et al. 2016) consider that there is a fourth approach commonly referred to as *embedded* because feature selection is achieved as part of a learning process, commonly through the optimization of a constrained regression model. However, in this survey, we will categorize them as filter multivariate, since in addition to jointly evaluate features, the primary objective of these methods is to perform feature selection (or ranking) rather than finding the classification or cluster labels. Moreover, we think that embedded methods could be considered a subcategory inside the main approaches (i.e., filter, wrapper, and hybrid), not hindering the possibility of having embedded methods in the three approaches.

Finally, using the information of Table 1, researchers of application areas could select the appropriate approach taking into consideration the characteristics of their data, their research objectives, and their specific constraints regarding runtime and expected quality.

5.2 Characteristics of feature selection methods for mixed data

In Table 2, we summarized the main characteristics of each feature selection method for mixed data reviewed in this survey. In this table, the first, second, and third columns show the name of the feature selection method, the type of method (S = Supervised or U = Unsupervised), and the approach (filter, wrapper, or hybrid), respectively. The fourth,

Table 2 Characteristics of the supervised and unsupervised feature selection methods for mixed data reviewed in this survey

FS method	Type	Approach	Apply feature transformations	Process features separately	Eliminate redundant features
CFS	S	Filter	✓	×	✓
FCBF	S	Filter	✓	×	✓
Relieff	S	Filter	×	×	×
mRMR	S	Filter	✓	×	✓
UFT	S	Filter	✓	×	✓
mDSM	S	Filter	✓	×	✓
Doan's et al.	S	Wrapper	✓	×	×
MFS	S	Hybrid	×	✓	×
RFE	S	Wrapper	×	×	×
ECMBF	S	Filter	×	×	✓
SSFSM	S	Filter	×	×	×
Mixed-MB	S	Filter	✓	×	✓
Hybrid MI	S	Hybrid	×	✓	×
Mixed MI	S	Hybrid	×	✓	×
HFS	S	Hybrid	×	×	✓
Coelho's et al.	S	Filter	×	×	×
RnR-SSFSM	S	Filter	✓	×	✓
FarVPN	S	Filter	×	×	✓
EFSH	S	Filter	×	×	×
Chen and Yang	S	Filter	×	×	×
Zhang's et al.	S	Hybrid	×	×	✓
FSMSD	S	Filter	×	✓	×
MIFSA/MIFSD	S	Filter	×	✓	×
SUD	U	Filter	✓	×	×
Dash and Liu's	U	Hybrid	✓	×	×
USFSM	U	Filter	×	×	×
Chaudhuri's	U	Hybrid	✓	×	✓
MOGA	U	Wrapper	×	×	×
CRAFT	U	Wrapper	×	×	×
VarSelLCM	U	Wrapper	×	×	×
DPM-MCMC	U	Wrapper	×	×	×

fifth, and last columns show an indicator (✓ = Yes or × = No) that points out if the method uses a feature transformation technique, processes numerical and non-numerical features separately, or can handle redundancy, respectively.

From Table 2, it is possible to observe that most feature selection methods for mixed data have been developed for the supervised case under the filter approach. Meanwhile, in the unsupervised case, wrapper methods have been the most explored. Likewise, we also observed that for the semi-supervised case (and as far as we know), no methods had been developed to address the feature selection problem in mixed data; therefore, this particular area represents a good research opportunity. On the other hand, we can observe

that most supervised methods consider eliminating redundant features; meanwhile, few methods address this problem for the unsupervised case. Finally, we have identified three main strategies commonly applied to address feature selection in mixed data: (1) applying feature transformation procedures, (2) handling features separately, and (3) defining and using different measures to evaluate features that can work with mixed data. In the following, we will describe these strategies in more detail and discuss their main advantages and drawbacks.

5.2.1 Applying feature transformations

A popular strategy for applying feature selection methods over mixed data is to perform feature transformation procedures, i.e., convert mixed-type data into single-type. The process of transforming a non-numerical feature into a numerical one is called *encoding*, and there are several methods in the literature (Breiman et al. 1984; Bruin 2011; Barcelo-Rico and Diez 2012; Cohen et al. 2013; Gniazdowski and Grabowski 2016). The main advantage of using these methods is that they provide a numerical representation for non-numerical features. Thus, the transformed data can be processed by algorithms developed to handle only numerical data. Nevertheless, it is well-known that encoding methods usually have the following drawbacks:

- Feature encoding introduces an artificial order between features values, which does not necessarily correspond to the original nature of the data (Gniazdowski and Grabowski 2016). The order induced by numerical values is commonly meaningless.
- Different relative distances are introduced that may not match the essence of the non-numerical data.
- A permutation of codes for non-numerical values can lead to different distance values (Doquire and Verleysen 2011a).
- Some mathematical operations such as addition and multiplication by a scalar do not make sense over the transformed data because they do not meet any algebraic, logical, or topological supposition on themselves (Ruiz-Shulcloper 2008).

Conversely, the process of transforming a set of numerical continuous features into non-numerical discrete ones through the association of categorical values to intervals is called *discretization*. Currently, there are a huge amount of discretization methods introduced in the literature (García et al. 2013), and they can be classified as either local or global. Local methods are characterized by operating on only one feature. Meanwhile, global methods consider all features (rather than one) before deciding where to induce interval breakpoints. Likewise, according to the availability of information, discretization methods can be categorized as supervised or unsupervised. Some representative supervised discretization methods, able to automatically determine the number of bins, are the Class-Attribute Interdependence Maximization (CAIM) (Kurgan and Cios 2004), ChiMerge (Kerber 1992), Chi2 (Liu and Setiono 1995), and Minimum Description Length Principle (MDLP) (Fayyad and Irani 1993). Likewise, some examples of unsupervised discretization methods are Equal frequency (Wong and Chiu 1987), Equal width (Wong and Chiu 1987), and Clustering-based discretizers (Chmielewski and Grzymala-Busse 1996; Monti and Cooper 1999; Hua and Zhao 2009). Feature discretization methods are widely used in practice since they allow taking advantage of those

methods designed to process categorical data. Nevertheless, it is well known that they have the following drawbacks:

- Discretization brings with it an inherent loss of information due to the binning process (Doquire and Verleysen 2011a; Foss et al. 2018).
- The results of the feature selection will highly depend on the applied discretization method.
- Some discretization methods (Cantú-Paz 2001; Hartemink and Gifford 2001; Dash et al. 2011) are sensitive to outliers.

5.2.2 Handling features separately

Another solution that has been considered, especially in some supervised feature selection methods (Tang and Mao 2005, 2007; Doquire and Verleysen 2011a, b; Kim and Jun 2018), is to analyze numerical and non-numerical features separately and then to join the two results. However, as De Leon and Chough (2013) pointed out, by using this solution, the associations that exist between numerical and non-numerical features are ignored. Moreover, many feature selection methods use distances/similarities among objects to find relevant features. However, if we split the features into numerical and non-numerical ones and compute the object distances/similarities separately, combining these two types of measures is non-trivial and not clear for the following reasons. (1) Both measures could calculate the object distance or similarity differently, which often turns out in non-comparable measures. (2) The scales of these measures could be non-similar. Therefore, the proportion in which two distance measures are combined is non-obvious.

5.2.3 Defining and using different measures to evaluate features in mixed data

The last strategy refers to those methods that use a same-principle-based measure computed differently for each feature type or those that use different measures to evaluate each feature type. In the former, the main drawback can be attributed to the nature of the criterion functions used. For example, most criterion functions used in information theory and fuzzy-rough set theory methods are based on probability distribution and granulation concepts, respectively. However, computing density estimations for numerical features in Information theory-based methods and determining a suitable scale parameter granulation for fuzzy-rough set theory methods is a challenging task (Li and Biswas 2002).

On the other hand, methods that use different measures to evaluate each feature type have the same problem of evaluating features separately; i.e., combining them does not turn out to be trivial. For example, some feature selection methods such as Paul and Dupont (2014), Paul et al. (2015), and Solorio-Fernández et al. (2017, 2019) use a similarity measure (kernels) (Wilson and Martinez 1997; Foss et al. 2018) that considers each type of feature differently using distinct feature evaluation measures. Nevertheless, combining these feature evaluation measures is not straightforward.

6 Concluding remarks, challenges and future research directions

With the diversification of data types, mixed data frequently occur in many applications, therefore developing feature selection methods that can handle such data has become an important and popular research topic. This paper presents a survey concerning feature selection methods for mixed data. Consequently, this work summarizes and discusses the state-of-the-art literature on supervised and unsupervised feature selection methods that can process mixed datasets. Moreover, the main advantages, disadvantages, challenges, and future directions for feature selection research are also discussed from the mixed data perspective. This survey aims to help researchers, practitioners, and academics to develop a better understanding of feature selection on mixed data, which we hope would help generate new ideas and new methods for solving feature selection on real-world problems.

Finally, based on our study and literature revision, we identify the following challenges and open problems on feature selection for mixed data.

- Most feature selection methods for mixed data have been introduced for supervised classification and just a few for unsupervised classification. Moreover, semi-supervised feature selection for mixed data has not been addressed. Hence, further developments on this last paradigm are necessary.
- In the unsupervised context, there are a considerable amount of clustering algorithms capable of handling mixed data (Balaji and Lavanya 2018; Ahmad and Khan 2019). However, unsupervised feature selection methods for this type of data are few. Therefore, more effort in this direction will undoubtedly benefit the research community since selecting a subset of relevant features can also enhance the interpretability of clustering algorithms.
- We note that feature transformation procedures such as encoding and discretization are very common in practice; however, as mentioned before, such procedures do not come without important disadvantages. Therefore, it is an open question to the research community to develop algorithms that can reduce feature transformation's adverse effects in feature selection for mixed data.
- Some feature selection methods for mixed data such as kernel and spectral-based rely on computing a similarity or distance matrix. However, this similarity matrix depends on a good definition of similarity or distance among objects. The notion of similarity is not clearly defined for mixed data (Dos Santos and Zárate 2015; Ahmad and Khan 2019), therefore performing feature selection remains challenging. Moreover, a detailed study is required to understand which similarity measures are more useful for feature selection in this context.
- As numerical and non-numerical data, real-world mixed data are imperfect, and missing values among features could occur. One plausible solution to this problem is to impute missing mixed data values and then perform feature selection. The other solution is to develop feature selection methods that can handle missing data in their objective function. However, the comparison between these two solutions has not been investigated in feature selection for mixed data, which could be interesting for the research community.
- In the supervised context, feature selection methods for mixed data commonly evaluate results regarding the quality of classification models. However, evaluating the performance of unsupervised feature selection methods is not straightforward.

Determining the best evaluation method remains an open problem, and more studies could be carried out in this direction.

- The suitable determination of the user-defined parameters represents a good opportunity for research. For example, the optimal number of features to select (in ranking-based methods), the number of clusters (in some unsupervised feature selection methods), and the determination of hyper-parameters inherent to each method are some of the open research challenges.
- In the Big Data era to deal with the processing of large volumes of data, the scalability and parallelization of feature selection methods represent a good area of opportunity for research and development (Li and Liu 2017). Hence, active research in this area is required to keep the field in synchronization with big data challenges. Similarly, developing fast and accurate online feature selection methods to handle large streams of mixed data requires attention to address shortcomings (Li et al. 2017).
- From our literature review, we have noticed that most of the reviewed works tested their methods on a few publicly available datasets. However, these datasets are relatively small in size and number of features and may not represent more complex problems. Therefore more studies in large and high dimensional datasets are needed. Moreover, we believe that sharing feature selection methods' implementations and real-world datasets in public domain tools would help compare and test them more fairly and realistically, undoubtedly benefiting the research community.

Acknowledgements The first author gratefully acknowledges to the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) for the collaboration grant awarded for the completion of this survey.

References

Aggarwal CC, Reddy CK (2013) Data clustering: algorithms and applications. CRC Press, Boca Raton

Ahmad A, Khan SS (2019) Survey of state-of-the-art mixed data clustering algorithms. IEEE Access 7:31883–31902

Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Selected papers of Hirotugu Akaike, pp 199–213. Springer

Alelyani S, Tang J, Liu H (2013) Feature selection for clustering: a review. In: Aggarwal CC, Reddy CK (eds) Data clustering: algorithms and applications, vol 29. CRC Press, Boca Raton, pp 110–121

Ang JC, Mirzal A, Haron H, Hamed HNA (2016) Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinf 13(5):971–989. https://doi.org/10.1109/TCBB.2015.2478454

Balaji K, Lavanya K (2018) Clustering algorithms for mixed datasets: a review. Int J Pure Appl Math 118(7):547–556

Barcelo-Rico F, Diez JL (2012) Geometrical codification for clustering mixed categorical and numerical databases. J Intell Inf Syst 39(1):167–185. https://doi.org/10.1007/s10844-011-0187-y

Ben Haj Kacem MA, Ben N'Cir CE, Essoussi N (2015) MapReduce-based k-prototypes clustering method for big data. In: Proceedings of the 2015 IEEE international conference on data science and advanced analytics, DSAA 2015 (October 2015), pp 4–6. https://doi.org/10.1109/DSAA.2015.7344894

Bharti KK, kumar Singh P (2014) A survey on filter techniques for feature selection in text mining. In: Proceedings of the second international conference on soft computing for problem solving (SocProS 2012), 28–30 Dec 2012, pp 1545–1559. Springer

Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2015) Feature selection for high-dimensional data. Springer, Berlin. https://doi.org/10.1007/978-3-319-21858-8

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth Inc, Belmont, CA

Bruin J (2011) newtest: command to compute new test {@ONLINE}. https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/
Cai J, Luo J, Wang S, Yang S (2018) Feature selection in machine learning: a new perspective. Neurocomputing. https://doi.org/10.1016/j.neucom.2017.11.077
Cantú-Paz E (2001) Supervised and unsupervised discretization methods for evolutionary algorithms. In: Workshop proceedings of the genetic and evolutionary computation conference (GECCO-2001), pp 213–216
Chandra B, Gupta M (2011) An efficient statistical feature selection approach for classification of gene expression data. J Biomed Inform 44(4):529–535. https://doi.org/10.1016/j.jbi.2011.01.001
Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024
Chaudhuri A, Samanta D, Sarma M (2021) Two-stage approach to feature set optimization for unsupervised dataset with heterogeneous attributes. Expert Syst Appl 172(January):114563. https://doi.org/10.1016/j.eswa.2021.114563
Chen D, Yang Y (2014) Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models. IEEE Trans Fuzzy Syst 22(5):1325–1334. https://doi.org/10.1109/TFUZZ.2013.2291570
Chen D, Zhang L, Zhao S, Hu Q, Zhu P (2012) A novel algorithm for finding reducts with fuzzy rough sets. IEEE Trans Fuzzy Syst 20(2):385–389. https://doi.org/10.1109/TFUZZ.2011.2173695
Chmielewski MR, Grzymala-Busse JW (1996) Global discretization of continuous attributes as preprocessing for machine learning. Int J Approx Reason 15(4):319–331
Coelho F, Braga AP, Verleysen M (2016) A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems. Int J Comput Intell Syst 9(4):726–733. https://doi.org/10.1080/18756891.2016.1204120
Cohen J, Cohen P, West SG, Aiken LS (2013) Applied multiple regression/correlation analysis for the behavioral sciences. Routledge, Abingdon
Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley-Interscience, Hoboken
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. Springer, Berlin, pp 110–121
Dash M, Liu H, Yao J (1997) Dimensionality reduction of unsupervised data. In: Proceedings ninth IEEE international conference on tools with artificial intelligence, pp 532–539. IEEE Computer Society. https://doi.org/10.1109/TAI.1997.632300. http://ieeexplore.ieee.org/document/632300/
Dash R, Paramguru RL, Dash R (2011) Comparative analysis of supervised and unsupervised discretization techniques. Int J Adv Sci Technol 2(3):29–37
De Leon AR, Chough KC (2013) Analysis of mixed data: methods & applications. CRC Press, Boca Raton
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM-Algorithm. JSTOR 39:1–22. https://doi.org/10.2307/2984875
Deng X, Li Y, Weng J, Zhang J (2019) Feature selection for text classification: a review. Multimedia Tools Appl 78(3):3797–3816
Devijver PA, Kittler J (1982) Pattern recognition: a statistical approach. Prentice Hall, Hoboken
Doan DM, Jeong DH, Ji SY (2020) Designing a feature selection technique for analyzing mixed data. In: 2020 10th annual computing and communication workshop and conference, CCWC 2020, Institute of Electrical and Electronics Engineers Inc., pp 46–52. https://doi.org/10.1109/CCWC47524.2020.9031193
Doquire G, Verleysen M (2011a) An hybrid approach to feature selection for mixed categorical and continuous data. In: Proceedings of the international conference on knowledge discovery and information retrieval, pp 394–401. https://doi.org/10.5220/0003634903940401. http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0003634903940401
Doquire G, Verleysen M (2011b) Mutual information based feature selection for mixed data. In: 19th European symposium on artificial neural networks, computational intelligence and machine learning (ESANN 2011), pp 333–338
Dos Santos TRL, Zárate LE (2015) Categorical data clustering: what similarity measure to recommend? Expert Syst Appl 42(3):1247–1260. https://doi.org/10.1016/j.eswa.2014.09.012
Dutta D, Dutta P, Sil J (2014) Simultaneous feature selection and clustering with mixed features by multi objective genetic algorithm. Int J Hybrid Intell Syst 11(1):41–54
Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5:845–889. https://doi.org/10.1016/j.patrec.2014.11.006
Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: IJCAI

Focant I, Hernandez-Lobato D, Ducreux J, Durez P, Toukap AN, Elewaut D, Houssiau FA, Dupont P, Lauwerys B (2011) Feasibility of a molecular diagnosis of arthritis based on the identification of specific transcriptomic profiles in knee synovial biopsies. Arthritis Rheum 63:abstract 1927:S751

Fop M, Murphy TB (2018) Variable selection methods for model-based clustering. Stat Surv 12:18–65. https://doi.org/10.1214/18-SS119

Foss AH, Markatou M, Ray B (2018) Distance metrics and clustering methods for mixed-type data. Int Stat Rev. https://doi.org/10.1111/insr.12274

Fowlkes EB, Gnanadesikan R, Kettenring JR (1988) Variable selection in clustering. J Classif 5(2):205–228. https://doi.org/10.1007/BF01897164

François D, Wertz V, Verleysen M (2006) The permutation test for feature selection by mutual information. In: ESANN 2006 Proceedings—European symposium on artificial neural networks, pp 239–244

García S, Luengo J, Sáez JA, López V, Herrera F (2013) A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Trans Knowl Data Eng 25(4):734–750. https://doi.org/10.1109/TKDE.2012.35

Garg VK, Rudin C, Jaakkola T (2016) CRAFT: ClusteR-specific Assorted Feature selecTion. In: Artificial intelligence and statistics, pp 305–313

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. J Am Stat Assoc 88(423):881–889

George EI, McCulloch RE (1997) Approaches for Bayesian variable selection. Stat Sin 7:339–373

Gniazdowski Z, Grabowski M (2016) Numerical coding of nominal data. arXiv preprint arXiv:1601.01966

Greco S, Matarazzo B, Slowinski R (2001) Rough sets theory for multicriteria decision analysis. Eur J Oper Res 129(1):1–47. https://doi.org/10.1016/S0377-2217(00)00167-3

Green PJ (1990) On use of the EM for penalized likelihood estimation. J R Stat Soc Ser B (Methodol) 52(3):443–452

Guyon I, Elisseeff A, De AM (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Hall MA (2000) Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the seventeenth international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '00, pp 359–366. http://dl.acm.org/citation.cfm?id=645529.657793

Hancer E, Xue B, Zhang M (2020) A survey on feature selection approaches for clustering. Artif Intell Rev 53(6):4519–4545. https://doi.org/10.1007/s10462-019-09800-w

Hartemink A, Gifford DK (2001) Principled computational methods for the validation and discovery of genetic regulatory networks. Massachusetts Institute of Technology. Ph.D. thesis, Ph. D. dissertation

He X, Cai D, Niyogi P (2005) Laplacian score for feature selection. In: Advances in neural information processing systems 18, vol 186, pp 507–514

Hennig C, Liao TF (2013) How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J R Stat Soc: Ser C: Appl Stat 62(3):309–369. https://doi.org/10.1111/j.1467-9876.2012.01066.x

Hu Q, Liu J, Yu D (2008a) Mixed feature selection based on granulation and approximation. Knowl-Based Syst 21(4):294–304

Hu Q, Yu D, Liu J, Wu C (2008b) Neighborhood rough set based heterogeneous feature subset selection. Inf Sci 178(18):3577–3594. https://doi.org/10.1016/j.ins.2008.05.024

Hua H, Zhao H (2009) A discretization algorithm of continuous attributes based on supervised clustering. In: 2009 Chinese conference on pattern recognition, pp 1–5. IEEE

Huang Z (1997) Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the 1st Pacific–Asia conference on knowledge discovery and data mining (PAKDD), Singapore, pp 21–34

Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Disc 2(3):283–304

Jensen R, Shen Q (2004) Fuzzy-rough attribute reduction with application to web categorization. Fuzzy Sets Syst 141(3):469–485. https://doi.org/10.1016/S0165-0114(03)00021-6

Jiang SY, Wang LX (2016) Efficient feature selection based on correlation measure between continuous and discrete features. Inf Process Lett 116(2):203–215. https://doi.org/10.1016/j.ipl.2015.07.005

Jović A, Brkić K, Bogunović N (2015a) A review of feature selection methods with applications. In: 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), pp 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Jović A, Brkić K, Bogunović N (2015b) A review of feature selection methods with applications. In: 2015 38th international convention on information and communication technology, electronics and microelectronics, MIPRO 2015—proceedings vol #, no May, pp 1200–1205, https://doi.org/10.1109/MIPRO.2015.7160458. http://ieeexplore.ieee.org/document/7160458/

Kerber R (1992) Chimerge: discretization of numeric attributes. In: Proceedings of the tenth national conference on Artificial intelligence, pp 123–128. Aaai Press
Kim KJ, Jun CH (2018) Rough set model based feature selection for mixed-type data with feature space decomposition. Expert Syst Appl 103:196–205. https://doi.org/10.1016/j.eswa.2018.03.010
Koller D, Sahami M (1996) Toward optimal feature selection. Technical report, Stanford InfoLab
Kononenko I (1994) Estimating attributes: analysis and extensions of RELIEF. In: Machine learning: ECML-94, pp 171–182. Springer
Kotsiantis SB (2011) Feature selection for machine learning classification problems: a recent overview. Artif Intell Rev 42:157–176. https://doi.org/10.1007/s10462-011-9230-1
Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E - Stat Nonlinear Soft Matter Phys 69(62):1–16. https://doi.org/10.1103/PhysRevE.69.066138
Kurgan LA, Cios KJ (2004) CAIM discretization algorithm. IEEE Trans Knowl Data Eng 16(2):145–153. https://doi.org/10.1109/TKDE.2004.1269594
Kwak N (2002) Input feature selection by mutual information based on Parzen window. IEEE Trans Pattern Anal Mach Intell 24(12):1667–1671. https://doi.org/10.1109/TPAMI.2002.1114861
Lam D, Wei M, Wunsch D (2015) Clustering data of mixed categorical and numerical type with unsupervised feature learning. IEEE Access 3:1605–1616. https://doi.org/10.1109/ACCESS.2015.2477216
Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, De Schaetzen V, Duque R, Bersini H, Nowé A (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinf 9(4):1106–1119. https://doi.org/10.1109/TCBB.2012.33
Lee J, Jeong JY, Jun CH (2020) Markov blanket-based universal feature selection for classification and regression of mixed-type data. Expert Syst Appl 158:113398. https://doi.org/10.1016/j.eswa.2020.113398
Lee PY, Loh WP, Chin JF (2017) Feature selection in multimedia: the state-of-the-art review. Image Vis Comput 67:29–42. https://doi.org/10.1016/j.imavis.2017.09.004
Li C, Biswas G (2002) Unsupervised learning with mixed numeric and nominal data. IEEE Trans Knowl Data Eng 14(4):673–690. https://doi.org/10.1109/TKDE.2002.1019208
Li J, Liu H (2017) Challenges of feature selection for big data analytics. IEEE Intell Syst 32(2):9–15. https://doi.org/10.1109/MIS.2017.38
Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2016) Feature selection: a data perspective. J Mach Learn Res:1–73. arXiv:1601.07996
Li Y, Li T, Liu H (2017) Recent advances in feature selection and its applications. Knowl Inf Syst 53(3):551–577. https://doi.org/10.1007/s10115-017-1059-8
Liang J, Zhao X, Li D, Cao F, Dang C (2012) Determining the number of clusters using information entropy for mixed data. Pattern Recogn 45(6):2251–2265. https://doi.org/10.1016/j.patcog.2011.12.017
Liang S, Ma A, Yang S, Wang Y, Ma Q (2018) A review of matched-pairs feature selection methods for gene expression data analysis. Comput Struct Biotechnol J 16:88–97. https://doi.org/10.1016/j.csbj.2018.02.005
Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml
Liu H, Motoda H (1998) Feature selection for knowledge discovery and data mining. Springer, Berlin. https://doi.org/10.1007/978-1-4615-5689-3
Liu H, Motoda H (2007) Computational methods of feature selection. CRC Press, Boca Raton
Liu H, Setiono R (1995) Chi2: feature selection and discretization of numeric attributes. In: TAI, p 388. IEEE
Liu H, Yu L, Member SS, Yu L, Member SS (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502. https://doi.org/10.1109/TKDE.2005.66
Liu H, Wei R, Jiang G (2013) A hybrid feature selection scheme for mixed attributes data. Comput Appl Math 32(1):145–161
Liu N (2012) The research of intrusion detection based on mixed clustering algorithm. In: Li Z, Li X, Liu Y, Cai Z (eds) Communications in computer and information science. CCIS, vol 316. Springer, Berlin, pp 92–100. https://doi.org/10.1007/978-3-642-34289-9_11
Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416. https://doi.org/10.1007/s11222-007-9033-z
Marbac M, Sedki M (2017) Variable selection for mixed data clustering: a model-based approach. eprint arXiv, arXiv:1703.02293
Marbac M, Sedki M (2019) VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values. Bioinformatics 35(7):1255–1257. https://doi.org/10.1093/bioinformatics/bty786

Miao J, Niu L (2016) A survey on feature selection. Procedia Comput Sci 91(Itqm):919–926. https://doi.org/10.1016/j.procs.2016.07.111

Mitra P, Murthy CA, Pal SK (2002) Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Mach Intell PAMI 24(3):301–312. https://doi.org/10.1109/34.990133

Monti S, Cooper GF (1999) A latent variable model for multivariate discretization. In: AISTATS

Mugunthadevi K, Punitha SC, Punithavalli M (2011) Survey on feature selection in document clustering. Int J Comput Sci Eng 3(3):1240–1244

Niu K, Niu Z, Su Y, Wang C, Lu H, Guan J (2015) A coupled user clustering algorithm based on mixed data for web-based learning systems. Math Probl Eng 2015:747628. https://doi.org/10.1155/2015/747628

Pal SK, Mitra P (2004) Pattern recognition algorithms for data mining, 1st edn. Chapman and Hall/CRC, London

Paul J, Dupont P (2014) Kernel methods for mixed feature selection. In: 22nd European symposium on artificial neural networks, computational intelligence and machine learning, ESANN 2014—proceedings, pp 301–306. Citeseer

Paul J, Dupont P (2015) Kernel methods for heterogeneous feature selection. Neurocomputing 169:187–195. https://doi.org/10.1016/j.neucom.2014.12.098

Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11(5):341–356. https://doi.org/10.1007/BF01001956

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Rao VM, Sastry VN (2012) Unsupervised feature ranking based on representation entropy. In: 2012 1st international conference on recent advances in information technology, RAIT-2012, pp 421–425. https://doi.org/10.1109/RAIT.2012.6194631

Remeseiro B, Bolon-Canedo V (2019) A review of feature selection methods in medical applications. Comput Biol Med 112(February):103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Ren M, Liu P, Wang Z, Lü L (2016) An improved kernel clustering algorithm for mixed-type data in network forensic. Int J Secur Appl 10(1):343–354. https://doi.org/10.14257/ijsia.2016.10.1.31

Rudnicki WR, Wrzesień M, Paja W (2013) Feature selection for data and pattern classification

Ruiz-Shulcloper J (2008) Pattern recognition with mixed and incomplete data. Pattern Recognit Image Anal 18(4):563–576. https://doi.org/10.1134/S1054661808040044

Ruiz-Shulcloper J, Abidi M (2002) Logical combinatorial pattern recognition: a review. Citeseer, pp 133–176

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Sang B, Chen H, Li T, Xu W, Yu H (2020) Incremental approaches for heterogeneous feature selection in dynamic ordered data. Inf Sci 541:475–501. https://doi.org/10.1016/j.ins.2020.06.051

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Sharmin S, Shoyaib M, Ali AA, Khan MAH, Chae O (2019) Simultaneous feature selection and discretization based on mutual information. Pattern Recognit 91:162–174. https://doi.org/10.1016/j.patcog.2019.02.016

Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MAZ (2017) A survey on semi-supervised feature selection methods. Pattern Recognit 64(November 2016):141–158. https://doi.org/10.1016/j.patcog.2016.11.003

Solorio-Fernández S, Martínez-Trinidad JF, Carrasco-Ochoa JA (2017) A new unsupervised spectral feature selection method for mixed data: a filter approach. Pattern Recognit 72:314–326. https://doi.org/10.1016/j.patcog.2017.07.020

Solorio-Fernández S, Martínez-Trinidad JF, Carrasco-Ochoa JA (2019) A supervised filter feature selection method for mixed data based on the spectral gap score. In: Carrasco-Ochoa JA, Martínez-Trinidad JF, Olvera-López JA, Salas J (eds) Pattern recognition. Springer International Publishing, Cham, pp 3–13

Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF (2020a) A review of unsupervised feature selection methods. Artif Intell Rev 53(2):907–948. https://doi.org/10.1007/s10462-019-09682-y

Solorio-Fernández S, Martínez-Trinidad JF, Carrasco-Ochoa JA (2020b) A supervised filter feature selection method for mixed data based on spectral feature selection and information-theory redundancy analysis. Pattern Recogn Lett 138:321–328. https://doi.org/10.1016/j.patrec.2020.07.039

Storlie CB, Myers SM, Katusic SK, Weaver AL, Voigt RG, Croarkin PE, Stoeckel RE, Port JD (2018) Clustering and variable selection in the presence of mixed variable types and missing data. Stat Med 37(19):2884–2899. https://doi.org/10.1002/sim.7697

Su X, Liu F (2018) A survey for study of feature selection based on mutual information. In: Workshop on hyperspectral image and signal processing, evolution in remote sensing, vol 2018-Sept, pp 1–4. https://doi.org/10.1109/WHISPERS.2018.8746913

Tang J, Alelyani S, Liu H (2014) Feature selection for classification: a review. In: Aggarwal CC (ed) Data classification: algorithms and applications. CRC Press, Boca Raton, p 37

Tang W, Mao K (2005) Feature selection algorithm for data with both nominal and continuous features. In: Ho TB, Cheung D, Liu H (eds) Advances in knowledge discovery and data mining: 9th Pacific–Asia conference, PAKDD 2005, Hanoi, Vietnam, 18–20 May 2005. Proceedings, pp 683–688. Springer, Berlin. https://doi.org/10.1007/11430919_78

Tang W, Mao KZ (2007) Feature selection algorithm for mixed data with both nominal and continuous features. Pattern Recognit Lett 28(5):563–571. https://doi.org/10.1016/j.patrec.2006.10.008

Tsamardinos I, Aliferis CF, Statnikov AR, Statnikov E (2003) Algorithms for large scale Markov blanket discovery. FLAIRS Conf 2:376–380

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536. https://doi.org/10.1038/415530a

Vergara JR, Estévez PA (2014) A review of feature selection methods based on mutual information. Neural Comput Appl 24(1):175–186. https://doi.org/10.1007/s00521-013-1368-0

Wang F, Liang J (2016) An efficient feature selection algorithm for hybrid data. Neurocomputing 193:33–41. https://doi.org/10.1016/j.neucom.2016.01.056

Wei M, Chow TWS, Chan RHM (2015) Heterogeneous feature subset selection using mutual information-based feature transformation. Neurocomputing 168:706–718. https://doi.org/10.1016/j.neucom.2015.05.053

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9(1):60–62. https://doi.org/10.1214/aoms/1177732360

Wilson DR, Martinez TR (1997) Improved heterogeneous distance functions. J Artif Intell Res 6:1–34. https://doi.org/10.1613/jair.346

Wong AKC, Chiu DKY (1987) Synthesizing statistical knowledge from incomplete mixed-mode data. IEEE Trans Pattern Anal Mach Intell PAMI PAMI–9(6):796–805. https://doi.org/10.1109/TPAMI.1987.4767986

Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, pp 267–273

Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp 856–863

Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224. https://doi.org/10.1145/1014052.1014149

Zhang R, Nie F, Li X, Wei X (2019) Feature selection with multi-view data: a survey. Inf Fusion 50:158–167. https://doi.org/10.1016/j.inffus.2018.11.019

Zhang X, Mei C, Chen D, Li J (2016) Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy. Pattern Recognit 56:1–15. https://doi.org/10.1016/j.patcog.2016.02.013

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

Zhao ZA, Liu H (2011) Spectral feature selection for data mining. Data mining and knowledge discovery series, 1st edn. Chapman & Hall/CRC, London. https://doi.org/10.1201/b11426

Zheng Z, Lei W, Huan L (2010) Efficient spectral feature selection with minimum redundancy. In: Twenty-fourth AAAI conference on artificial intelligence, pp 1–6