



**I  
N  
A  
O  
E**

# **Unsupervised Feature Selection Method for Mixed Data**

Saúl Solorio-Fernández, Ariel Carrasco-Ochoa

Technical Report No. CCC-19-005  
November 2019

© Computer Science Department  
INAOE

Luis Enrique Erro 1  
Sta. Ma. Tonantzintla,  
72840, Puebla, México.



# Unsupervised Feature Selection Method for Mixed Data

Saúl Solorio Fernández

Jesús Ariel Carrasco Ochoa

Computer Science Department

National Institute of Astrophysics, Optics and Electronics

Luis Enrique Erro # 1, Santa María Tonantzintla, Puebla, 72840, México

E-mail: {sausolofer, ariel}@inaoep.mx

## Abstract

*In recent years, unsupervised feature selection methods have attracted considerable interest in different areas; this is mainly due to their ability to identify and remove irrelevant and/or redundant features without needing a supervised dataset. However, most of these methods can only process numerical data; so in practical problems in areas such as medicine, economy, business, and social sciences, where it is common that objects are described by numerical and non-numerical features (mixed data), these methods cannot be directly applied. To overcome this limitation, in practice, it is common to apply an encoding method over non-numerical features. Nevertheless, in general, this approach is not a good choice, since by coding data we incorporate a notion of order into the feature values that does not necessarily correspond to the nature of the original dataset. Moreover, the permutation of codes for two values can lead to different distance values, and some mathematical operations do not make sense over the transformed data. For this reason, this Ph.D. research proposal focuses on developing a new unsupervised feature selection method for mixed datasets.*

**Keywords**— Feature selection, Unsupervised feature selection, Mixed data

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Theoretical Basis</b>	<b>5</b>
2.1	Unsupervised Feature Selection . . . . .	5
2.2	Clustering algorithms for mixed data . . . . .	8
2.3	Performance Evaluation . . . . .	10
<b>3</b>	<b>Related Work</b>	<b>11</b>
3.1	Filter approach . . . . .	11
3.2	Wrapper approach . . . . .	14
3.3	Hybrid approach . . . . .	15
3.4	Concluding remarks . . . . .	16
<b>4</b>	<b>Justification and Motivation</b>	<b>17</b>
<b>5</b>	<b>Ph.D. Research Proposal</b>	<b>17</b>
5.1	Problem to be solved . . . . .	17
5.2	Research questions . . . . .	17
5.3	Hypothesis . . . . .	17
5.4	Objectives . . . . .	18
5.5	Methodology . . . . .	18
5.6	Expected contributions . . . . .	20
5.7	Publication plan . . . . .	20
5.8	Schedule . . . . .	22
<b>6</b>	<b>Preliminary Results</b>	<b>23</b>
6.1	A new filter unsupervised spectral feature selection method for mixed data . . . . .	23
6.2	Experimental results . . . . .	26
6.3	Discussion . . . . .	40
<b>7</b>	<b>Conclusions</b>	<b>41</b>

## 1 Introduction

Feature selection [1–4] (also known as attribute selection) appears in different areas such as pattern recognition, machine learning, data mining and multivariate statistical analysis. In these areas often the objects<sup>1</sup> of study may include in their description irrelevant and redundant features [5] that can significantly affect the analysis of the data, resulting in biased or even incorrect models [6]. Feature selection methods aim to select only those features that allow improving the classification results. Moreover, feature selection does not only reduce the dimensionality of the data; but it also leads to more compact models and possibly better generalization ability [7].

Over the last decades, many feature selection methods have been proposed, most of them for supervised problems [8–12]. However, there are many real applications where unsupervised problems arise (i.e., the objects are not labeled) [13, 14], and supervised features selection methods cannot be applied. Under this context, unsupervised feature selection [15, 16] has gained significant interest into the scientific community [17] since according to [1, 18, 19], unsupervised feature selection methods have two important advantages. 1) they are unbiased and perform well when prior knowledge is not available, and 2) they can reduce the risk of data over-fitting in contrast to supervised feature selection that may be unable to deal with a new class of data.

Several recent works [20–28] show that unsupervised feature selection is an active research area, with applications in genomic analysis [13, 29, 30], text mining [14, 31, 32], image retrieval [33–35], and intrusion detection [36, 37], to name a few. However, most unsupervised feature selection methods developed so far, are exclusively for numerical data; therefore they are not directly applicable on datasets where the objects are described by both numerical and non-numerical features (mixed data). Mixed data [38, 39] are very common, and they appear in many problems. For example, in biomedical and health-care applications [40–42], socioeconomics and business [38], software cost estimations [43], etc.

In practice, for applying unsupervised feature selection methods for numerical datasets over mixed data, is common to perform a previous process of data transformation. The process of transforming from a non-numerical feature to a numerical feature is called encoding, and there are several methods for achieving this, such as ordinal coding, one-hot, binary, polynomial, etc. [44–46]. All of them aim the same purpose; providing a numerical meaning for non-numerical features in such a way they can be processed by algorithms developed to handle numerical data. However, whatever the encoding method used, this process has the following disadvantages:

1. Feature codification introduces an artificial order between features values, which does not necessarily correspond to the original nature of the dataset [44].
2. Different relative distances are introduced that may not match the essence of the data.
3. The permutation of codes for two non-numerical values can lead to different distance values [47].
4. 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 [48].

On the other hand, as we will see later in the related work section, there are some unsupervised feature selection methods that perform an a priori discretization for transforming numerical features into non-numerical ones. Nevertheless, this discretization brings with it an inherent loss of information due to the

---

<sup>1</sup>Also called instances, records, observations or samples. Throughout this research proposal, we will refer to them as objects.

binning process [47], and the results of feature selection will highly depend on the applied discretization method. Additionally, it is known that some unsupervised discretization methods are sensitive to outliers [49–51].

As far as we know, in the literature of unsupervised feature selection, only three works claim they can be applied over mixed datasets. These works will be discussed in detail in the related work section. However, it is important to highlight that two of them make feature transformation before performing feature selection. Regarding the third one, the authors introduced a data clustering method, and during the clustering process, this method performs feature selection. Moreover, as we will explain later, this method requires the setting of several parameters, and it is computationally expensive.

In view of the above mentioned, in this Ph.D. research proposal the unsupervised feature selection problem over mixed datasets is addressed, i.e., on datasets where the objects are described simultaneously by numerical and non-numerical features.

The rest of this document is organized as follows: In section 2, theoretical basis about unsupervised feature selection are given. Section 3 reviews the related work, and in section 4, the justification and motivation for the development of this research are presented. The Ph.D. research proposal is presented in section 5, including a description of the problem to be solved, research questions, hypothesis, objectives, methodology, expected contributions, publication plan and the schedule of activities. Preliminary results are presented in section 6, and finally, in section 7 our conclusions are provided.

## 2 Theoretical Basis

In this section, the necessary background of the main topics required for understanding the contents of this research is presented. First, we define the unsupervised feature selection problem. Then, the main approaches in unsupervised feature selection are presented. After, a brief description of the most popular clustering algorithms for mixed data is given. Finally, we discuss the measures and strategies commonly used to assess unsupervised feature selection methods.

### 2.1 Unsupervised Feature Selection

In the supervised classification, feature selection methods maximize some objective function linked to class prediction. In this context, since class labels are available, it is natural to keep only the features that are related to or lead to these classes. However, in unsupervised classification, we are not given class labels, in fact, the goal is to find groups (also known as clusters). Therefore, the following questions arise: which features should we keep? Why not use all the information that we have? The problem is that not all the features are important or relevant. Some of the features may be redundant, and some others may be irrelevant [5, 16].

According to [4], a relevant feature (also known as a consistent feature) contains information about the objective concept<sup>2</sup>, and therefore, in unsupervised classification, it helps to find good cluster structures in the data. Conversely, an irrelevant feature does not allow to distinguish good cluster structures in the data. To clarify this, in Figure 1a an example of relevant and irrelevant features is shown. In this figure, we can see that  $F_1$  is a relevant feature because it can separate the data; as it can be seen when the data are projected to its respective axis. On the contrary,  $F_2$  is an example of an irrelevant feature because it by itself is unable to separate the data, as we can see in its projection.

---

<sup>2</sup> In the case of unsupervised (clustering) tasks, this concept is closely linked to those features that reveal interesting and natural structures underlying the data [16].

On the other hand, a redundant feature refers to a feature that is relevant for discovering cluster structures in the data. But if it is removed from the data, it has not negative effect due to the existence of another feature (or set of features) that provides the same information (see Figure 1b). Redundant features unnecessarily increase the dimensionality, and therefore they can be removed [5]. Furthermore, it has been empirically shown that removing redundant features can result in significant performance improvement [9].

### 2.1.1 Formal problem statement

According to [52–54], a good feature selection method should select a subset of features that are not only individually relevant but also with low redundancy. Therefore, the unsupervised feature selection problem can be formulated as follows:

*Given a collection of  $m$  objects  $X = \{x_1, x_2, \dots, x_m\}$ , described by a set of  $n$  features  $T = \{F_1, F_2, \dots, F_n\}$  possibly of different type (mixed data). Unsupervised feature selection consists in identifying a subset of features  $T' \subseteq T$ , without using class label information, such that  $T'$  does not contain irrelevant and/or redundant features, and good cluster structures in the data can be obtained or discovered.*

### 2.1.2 Approaches

Similar to feature selection for supervised classification, unsupervised feature selection methods can be categorized in three main approaches [55]: filter, wrapper, and hybrid.

Methods based on the filter approach select features based only on the intrinsic properties of the data, such as the variance of features, similarity among features, consistency, entropy, etc. A typical filter method is composed by two components (see Figure 2); 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 some intrinsic quality measure. This process ends until some pre-established stop criterion is met. As it can be seen, these

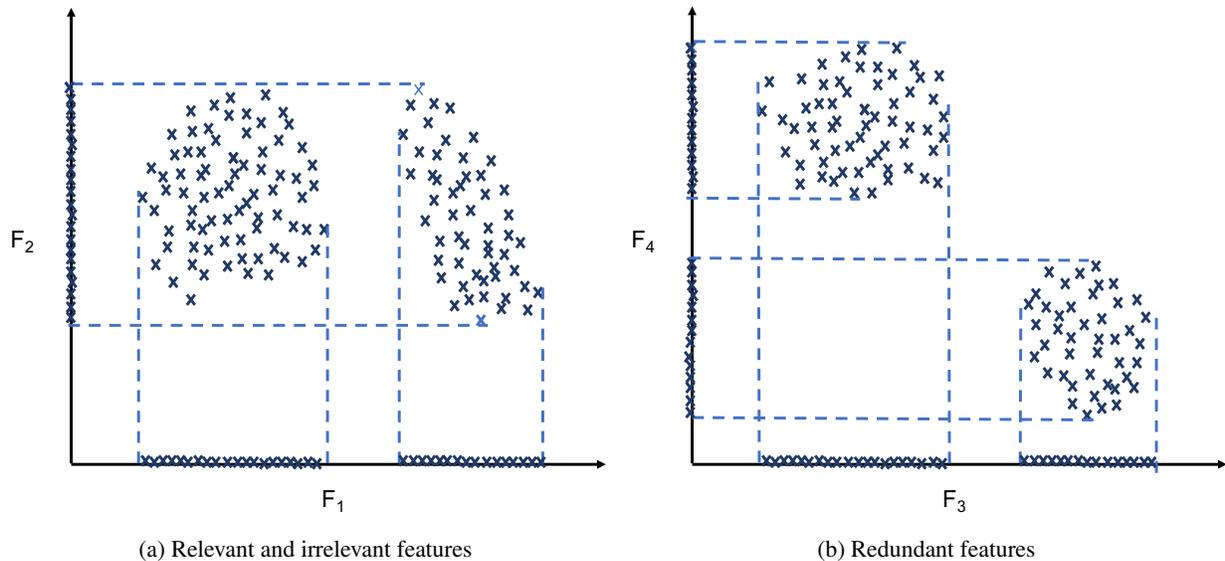


Figure 1: Relevant, irrelevant and redundant features.

methods are very simple, and unlike methods based on the wrapper approach they do not need a clustering algorithm for finding relevant features; consequently they tend to be faster and scalable.

According to [55], filter methods can be classified as univariate and multivariate. Univariate methods evaluate each feature in order to get an ordered list (ranking). These methods can effectively identify and remove irrelevant features, but they are unable to remove redundant ones because they do not take into account possible dependencies between features. On the other hand, filter multivariate methods assume that the dependent features should be discarded, being the independent ones the most relevant. These later methods can handle redundant and irrelevant features.

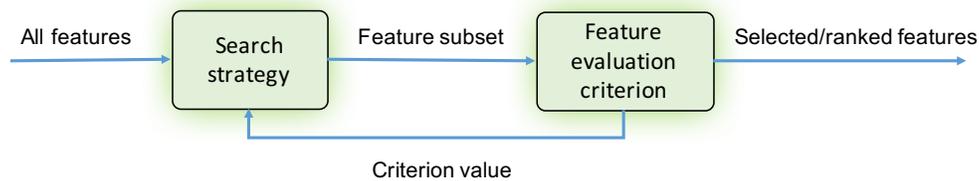


Figure 2: Filter approach for unsupervised feature selection.

On the other hand, methods based on the wrapper approach evaluate feature subsets depending on their performance under a specific clustering algorithm. As it is shown in Figure 3, a typical unsupervised feature selection method based on the wrapper approach consists of three basic components, namely, a search strategy, a clustering algorithm, 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 clustering algorithm is applied to the data described by the candidate feature subset. In the final component, clusters are evaluated according to a feature evaluation criterion. The subset that best fits the evaluation criterion will be chosen from all the candidates that have been evaluated. These methods usually obtain good feature subsets. However, the main disadvantages of these methods are that they usually have a high computational cost, and they have to be applied in conjunction with a particular clustering algorithm.

According to [2], methods based on the wrapper approach can be divided into two broad categories; sequential and bio-inspired. In the former, features are added or removed sequentially, but they tend to be trapped in local optimal solutions. The latter, on the other hand, try to incorporate randomness into their search procedure for escaping from these local solutions.

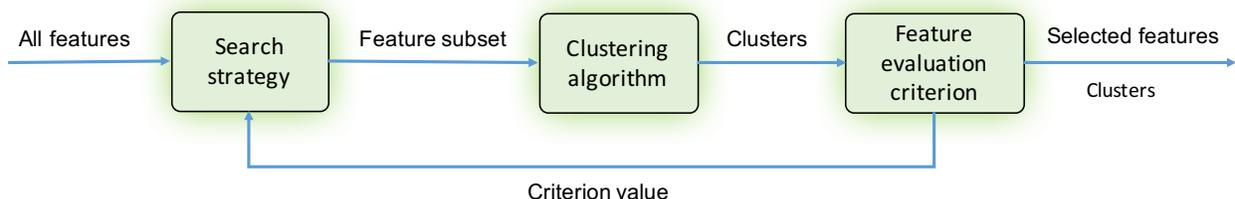


Figure 3: Wrapper approach for unsupervised feature selection.

Finally, methods based on the hybrid approach try to exploit the advantages of both, filter and wrapper, approaches, with the aim of obtaining a reasonable compromise between efficiency (computational effort) and effectiveness (good feature subsets). A typical hybrid method goes through the following steps: (1) it

employs a filter criterion to select different candidate subsets or to produce a feature ranking. Then, (2) it evaluates the quality of clustering for each candidate subset or some feature subsets based on the feature ranking. In the final step, (3) the subset with the highest clustering quality is selected. Algorithms belonging to the hybrid approach usually produce better clustering quality than those of the filter approach, but, they are less efficient. Compared to the wrapper approach, the hybrid methods are much more efficient [56], but they usually produce lower quality solutions.

## 2.2 Clustering algorithms for mixed data

Given a data sample, the objective of clustering is to group similar objects together. Generally, the clustering algorithms can be classified as partitional and hierarchical. Partitional clustering provides one level of clustering. Hierarchical clustering, on the other hand, provides multiple levels (a hierarchy) of clustering solutions. There are several algorithms for performing clustering, a survey of these algorithms can be found in [57, 58].

Since one of the tasks involved in this research proposal is the clustering of mixed data, in this section, we briefly present two popular partitional clustering algorithms for this kind of data:  $k$ -prototypes and finite mixture model clustering.

### 2.2.1 $k$ -prototypes

The  $k$ -prototypes algorithm [59] is a clustering algorithm designed to cluster mixed datasets. This algorithm is based on the idea of  $k$ -means [60].

Let  $X$  be a mixed dataset containing  $m$  objects, each one described by  $n$  features. Without loss of generality, we assume that the first  $p$  features are numerical and the last  $n - p$  features are non-numerical. The distance between two objects  $\mathbf{x}$  and  $\mathbf{y}$  in  $X$  can be defined as [59]:

$$D_{mix}(\mathbf{x}, \mathbf{y}, \beta) = \sum_{i=1}^p (x_i - y_i)^2 + \beta \sum_{i=p+1}^n \delta(x_i, y_i) \quad (1)$$

where  $x_i$  and  $y_i$  are the  $i^{th}$  component of  $\mathbf{x}$  and  $\mathbf{y}$  respectively,  $\beta$  is a balance weight used to avoid favoring either type of attribute, and  $\delta(\cdot, \cdot)$  is the simple matching distance defined as

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases}$$

The objective function that  $k$ -prototypes tries to minimize is defined as

$$P_\beta = \sum_{j=1}^k \sum_{\mathbf{x} \in C_j} D_{mix}(\mathbf{x}, \boldsymbol{\mu}_j, \beta) \quad (2)$$

where  $D_{mix}(\cdot, \cdot, \beta)$  is defined in (1),  $k$  is the number of clusters,  $C_j$  is the  $j^{th}$  cluster, and  $\boldsymbol{\mu}_j$  is the center or prototype of the cluster  $C_j$ .

To minimize the objective function defined in (2), the algorithm iteratively updates the cluster memberships given the cluster centers, and updates the cluster centers given the cluster memberships until a stop condition is met.

At the beginning, the  $k$ -prototypes algorithm initializes  $k$  cluster centers by randomly selecting  $k$  distinct objects from the dataset. Suppose  $\mu_1^{(0)}, \mu_2^{(0)}, \dots, \mu_k^{(0)}$  are the  $k$  initial cluster centers. The  $k$ -prototypes algorithm updates the cluster memberships  $\gamma_1, \gamma_2, \dots, \gamma_m$  as follows:

$$\gamma_i^0 = \operatorname{argmin}_{1 \leq j \leq k} D_{mix}(\mathbf{x}_i, \mu_j^{(0)}, \beta), \quad (3)$$

where  $D_{mix}(\cdot, \cdot, \beta)$  is defined in (1).

Once the cluster memberships have been updated, the algorithm proceeds to update the cluster centers as follows:

$$\begin{aligned} \mu_{jh}^{(1)} &= \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} x_h, \quad h = 1, 2, \dots, p \\ \mu_{jh}^{(1)} &= \operatorname{mode}_h(C_j), \quad h = p + 1, p + 2, \dots, n, \end{aligned} \quad (4)$$

where  $C_j = \{\mathbf{x}_i \in X : \gamma_i^{(0)} = j\}$  for  $j = 0, 1, \dots, k$ , and  $\operatorname{mode}_h(C_j)$  is the most frequent non-numerical value of the  $h^{\text{th}}$  feature in cluster  $C_j$ . The  $k$  prototypes algorithm repeats the above steps until the cluster memberships do not change or the maximum number of iterations is reached.

### 2.2.2 Finite mixture models

A finite mixture model assumes that a dataset  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$  with  $m$  objects is generated from a mixture of  $k$  component density functions, in which  $p(\mathbf{x}_i|\theta_j)$  represents the density function of component  $j$  for  $j = 1, 2, \dots, k$ , where  $\theta_j$  is the parameter (to be estimated) for cluster  $j$ . The probability density function for an object  $\mathbf{x}_i$  is expressed by

$$p(\mathbf{x}_i) = \sum_{j=1}^k \alpha_j p(\mathbf{x}_i|\theta_j) \quad (5)$$

where the  $\alpha_j$  are the mixing proportions of the components (subject to  $\alpha_j \geq 0$  and  $\sum_{j=1}^k \alpha_j = 1$ ). The log-likelihood of  $X$  is then given by

$$\mathcal{L}(\Theta|X) = \sum_{i=1}^m \ln \sum_{j=1}^k p(\mathbf{x}_i) \quad (6)$$

where  $\Theta$  is the set containing the mixture parameters  $\theta_j$ . For optimizing (6), it is necessary to estimate  $\Theta$ . This optimization is commonly achieved applying the Expectation-Maximization (EM) [61] algorithm which allows finding a (local) maximum likelihood or maximum a posteriori (MAP) estimate of the parameters for the given dataset. The EM algorithm iterates between an Expectation-step (E-step), which computes the expected complete data log-likelihood given the observed data and the model parameters, and a Maximization-step (M-step), which estimates the model parameters by maximizing the expected complete data log-likelihood from the E-step, until convergence. The clustering solution that we obtain in a mixture model is “soft” because we obtain an estimated cluster membership (i.e., each object belongs to all clusters with some probability of belonging to each cluster), in contrast to  $k$ -prototypes which provides a “hard”-clustering solution (i.e., each object only belongs to a single cluster).

Mixture-based models can deal with different types of features [4]. A Gaussian distribution is typically used for numerical features and multinomials for non-numerical features (Gaussian-multinomial mixture)

[62]. A more thorough description of clustering using Gaussian-multinomial finite mixture models can be found in [63–66].

### 2.3 Performance Evaluation

There are two standard ways to assess the performance of unsupervised feature selection methods [17]: in terms of clustering results and in terms of supervised classification results. For assessing clustering results, the clustering accuracy (ACC) and the Normalized Mutual Information (NMI) evaluation measures are commonly used. These evaluation measures are defined as follows:

Denoting by  $q_i$  the label computed by a clustering algorithm and by  $p_i$  the true label of  $x_i$ . ACC [67] is computed as follows:

$$ACC = \frac{\sum_{i=1}^m \delta(p_i, \text{map}(q_i))}{m} \quad (7)$$

where  $m$  is the total number of objects and  $\delta(x, y) = 1$  if  $x = y$ ; otherwise  $\delta(x, y) = 0$ .  $\text{map}(q_i)$  is a mapping function that permutes clustering labels to get the best match with the true labels. ACC values range from 0 to 1 where larger ACC means a better clustering.

Given the clustering results and the true labels  $P$  and  $Q$ , respectively, NMI [67] is defined as

$$NMI = \frac{I(P, Q)}{\max\{H(P), H(Q)\}} \quad (8)$$

where  $H(P)$  and  $H(Q)$  are the entropies of  $P$  and  $Q$  respectively, and  $I(P, Q)$  is the mutual information [68] between  $P$  and  $Q$  defined as follows:

$$I(P, Q) = \sum_{p_i \in P, q_j \in Q} p(p_i, q_j) \cdot \log_2 \frac{p(p_i, q_j)}{p(p_i) \cdot p(q_j)}$$

where  $p(p_i)$  and  $p(q_j)$  are the probabilities that an object arbitrarily selected from the dataset belongs to the clusters  $p_i$  and  $q_j$ , respectively, and  $p(p_i, q_j)$  is the joint probability that an arbitrarily selected object belongs to the clusters  $p_i$  and  $q_j$  at the same time. Values of NMI range from 0 to 1. NMI reflects how identical or independent are  $P$  and  $Q$ . Hence, better clusterings get larger NMI values.

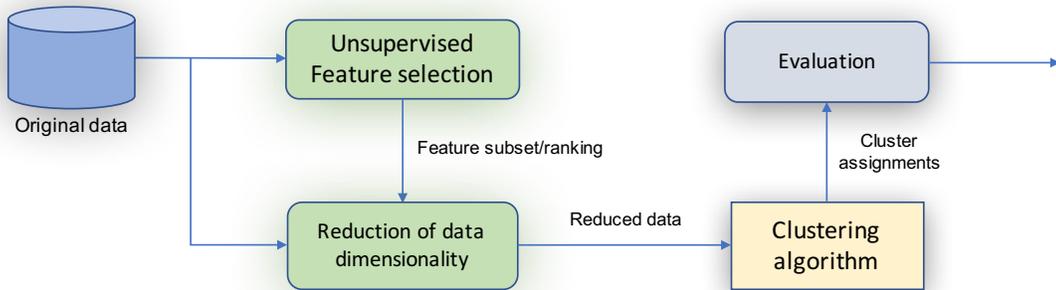


Figure 4: Feature selection evaluation using the clustering results.

To evaluate the performance of unsupervised feature selection methods in terms of clustering results, the steps showed in Figure 4 are commonly followed. An unsupervised feature selection method is first applied over the original dataset to select/rank features. Then, a clustering algorithm is applied over the reduced data, and the clustering results are assessed through an evaluation measure. In these last two steps, because in our experiments we will use  $k$ -prototypes and EM as clustering algorithms, and since these depend on the initial centers, we repeat these algorithms  $k$  times with different initialization points and the average of these results is reported.

On the other hand, for assessing the results of unsupervised feature selection methods regarding supervised classification results, the accuracy ACC (or error rate) of a classifier is commonly used. This evaluation is performed as follows: the whole dataset is usually divided into two parts - a training set and a test set. An unsupervised feature selection method is first applied on the training set to obtain a subset of relevant features, but without using the class labels. Then 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. These evaluation steps are shown in Figure 5.

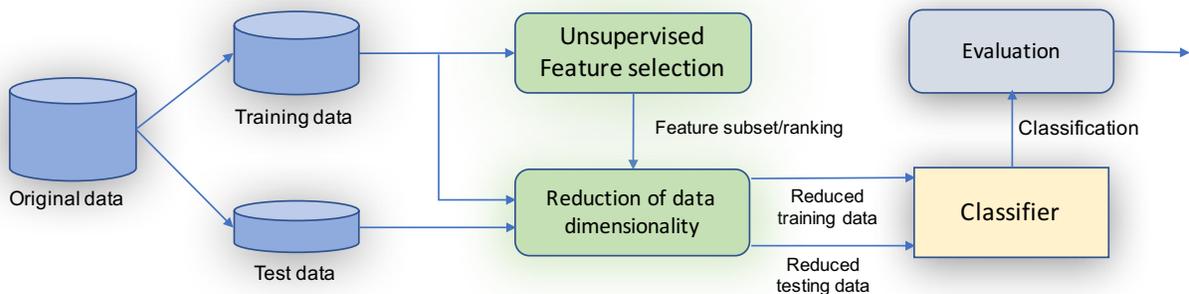


Figure 5: Feature selection evaluation using the classification results.

### 3 Related Work

In this section, we review the main unsupervised feature selection methods (filter, wrapper, and hybrid) reported in the literature. For this, we will follow the taxonomy shown in Figure 6.

#### 3.1 Filter approach

##### 3.1.1 Univariate

One of the first works developed in this category was introduced by Dash et al. in [69]. In this work, the authors introduced a new filter method called Sequential backward selection method for Unsupervised Data (SUD). This filter-based method weighs features using a measure of “entropy of similarities”, which is defined as the total entropy induced from a similarity matrix  $W$ , where the elements of  $W$  represent the similarity of each pair of objects in the dataset. The similarities in  $W$  are calculated as follows: when

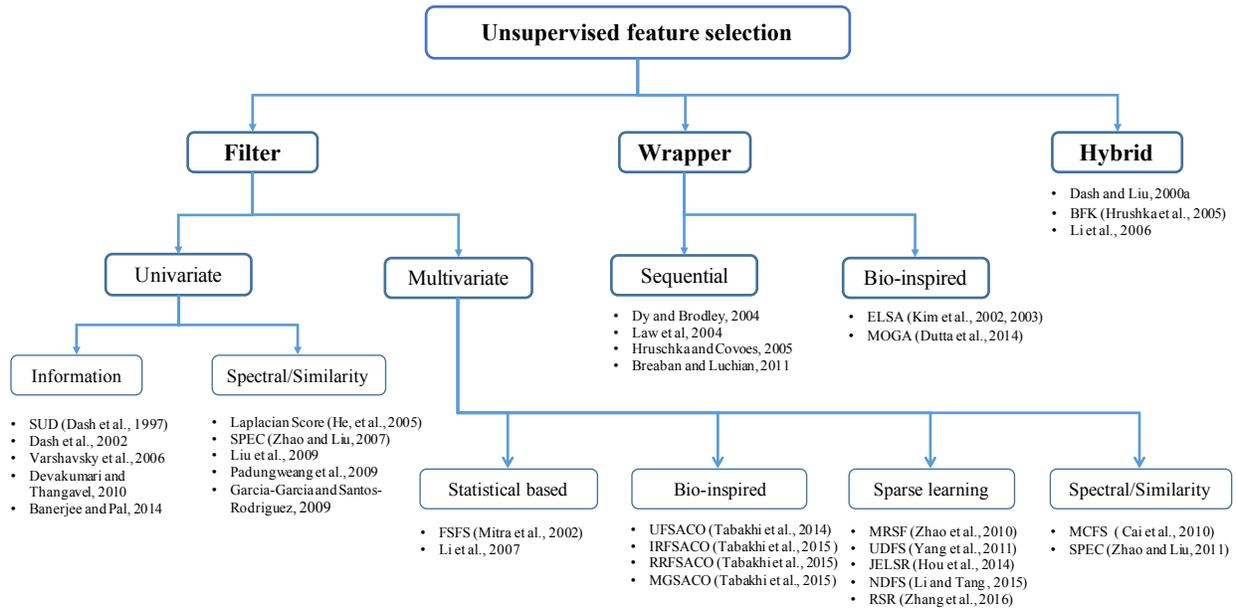


Figure 6: Taxonomy of unsupervised feature selection methods.

all features describing the objects in the dataset are numerical, the similarities are calculated using the exponential of the Euclidean distance metric; on the other hand, when the whole features in the dataset are non-numerical, the Hamming distance is used. For mixed data, the authors recommend performing a feature discretization before using the Hamming distance. The relevance of each feature is quantified using a leave-one-out sequential backward strategy jointly with the entropy measure above mentioned. The final result is a feature ranking ordered from the most to the least relevant feature. A later work based on this same idea was introduced in [70], where the main difference regarding the first one, is that instead of getting a feature ranking, a subset of features is selected using a forward selection search. Furthermore, the authors do not mention that their method can be applied to mixed data.

On the other hand, in [71] an unsupervised feature selection method for numerical data called SVD-Entropy is proposed. This method uses a measure based on the Singular Value Decomposition (SVD) of the data matrix [72]. The basic idea of this method is to measure the entropy of the data according to their singular values. When the entropy is low well-formed clusters are generated; by contrast, when the entropy is high the spectrum is uniformly distributed. In this work, some strategies including simple ranking, forward selection, and backward elimination were used. Two most recent works based on this measure of entropy were introduced in [19, 73]. In these last works, it is tried to avoid some drawbacks of SVD-Entropy, such as the weaknesses of forward selection and backward elimination searches, and the inability to distinguish features having a constant value.

Another relatively recent unsupervised feature selection methods for numerical data that has won acceptance for its affectivity, robustness and scalability are those methods based on spectral feature selection [6]. Spectral feature selection is based on the spectral graph theory [74], linear algebra and mathematical optimization. The general idea of spectral feature selection methods is to construct an affinity matrix  $W$  from the data similarities. This matrix represents the local or global data structure depending on the number of neighbors specified. Subsequently, from the affinity matrix, Laplacian or normalized Laplacian matrices are generated, which have many useful properties for feature selection. The Laplacian matrices or their

eigen-system associated are used by score functions to measure the relevance/consistency of each feature. Examples of univariate methods based on this approach are SPEC [75], Laplacian Score [67] and its derivative methods [76–78]. Spectral feature selection has been also used in multivariate methods; some examples are given in the following sub-section.

### 3.1.2 Multivariate

One of the most representative and referenced works in this category is FSFS (Feature Selection using Feature Similarity) introduced by Pabitra Mitra et al. in [79]. In this work, the authors introduced a measure of dependency/similarity to reduce feature redundancy; this measure called Maximal Information Compression Index (MICI) is based on the variance-covariance between features. The method involves partitioning the original set of features into clusters or groups, such that those features in the same cluster are highly similar (using MICI), while those in different clusters are dissimilar. Partitioning of the features is done based on the KNN principle. In the last stage, once the clusters are formed, FSFS selects only one feature from each cluster to form the final feature subset. Likewise, following a similar idea, in [80] a hierarchical method that tries to remove both redundant and irrelevant features is proposed. This method uses the MICI index proposed in [79] to remove redundant features. Subsequently, an exponential entropy measure is used to sort features according to their relevance. At the end, a non-redundant feature subset is selected using the fuzzy evaluation index FFEI [81].

Recently several bio-inspired unsupervised feature selection methods based on swarm intelligence framework [82, 83] have been proposed. In [84], a method based on this idea called Unsupervised Feature Selection based on Ant Colony Optimization (UFSACO) is proposed, whose main objective is to select feature subsets with low redundancy. In this work, first of all, the search space is represented as an undirected graph completely connected; where the set of nodes represent each feature and the weight of each edge denotes the similarity between nodes. This similarity is calculated using cosine similarity function, where the authors follow the idea that if two features resemble this similarity, then the features are redundant. Each node in the graph has a desirability measure associated called pheromone, which is updated by agents (ants) in function of its current value, a pre-specified decay rate, and the number of times that a given feature is selected by any agent. The complete procedure is performed as follows: at the beginning, a constant amount of initial pheromone to each node is assigned, then, the agents randomly placed in each node traverse the graph according to a transition rule (greedy and probabilistic) and a maximum number of nodes to visit. The transition rule that moves an agent from a node  $u$  to the other  $v$  is in function of the pheromone value of  $v$  and the inverse of the similarity measure between  $u$  and  $v$ , so the agents prefer high pheromone values and low similarities. The agents traverse the graph iteratively following the transition rule above mentioned until a pre-specified stop criterion (number of iterations) is reached. Finally, those features with the highest pheromone value are selected. Thus, it is expected to select feature subsets with low redundancy. Other three methods based on this idea are MGSACO [85], IRFSACO, and RRFSACO [54]. The difference is that these methods while minimizing redundancy between features, also aim to maximize the relevance.

Other multivariate feature selection methods that have received much attention in the last years due to their good performance and interpretability [17] are those based on sparse learning. Sparse learning [86] refers to that collection of learning methods that seek a trade-off between some goodness-of-fit measure and sparsity of the result, the latter property is used in many supervised and unsupervised feature selection methods. The general idea of feature selection methods based on sparse learning is to minimize fitting errors along with some sparse regularization terms. The sparse regularization forces some feature coefficients to be small or exactly zero; then the corresponding features can be simply eliminated. Examples of methods

based on this idea are: RSR [87], UDFS [88] NDFS [89, 90], JELSR [91, 92] and MRSF [93].

Finally, some methods based on multivariate spectral feature selection combined with sparse learning have also been developed, an example of these methods is introduced by Cai et al. in [94], where an unsupervised feature selection method to measure the correlation between features called Multi-Cluster Feature Selection (MCFS) is proposed. MCFS consists of three steps: (1) the spectral clustering step, (2) sparse coefficient learning step and (3) feature selection step. In the first step, spectral clustering is applied on the dataset to detect the multi-cluster structure of the data. In the second step, since the embedding of the data is known, through of the first  $k$  eigenvectors of the Laplacian matrix MCFS measures the importance of a feature by a regression model with a  $L_1$ -norm regularization [95]. In the third step, after solving the regression problem, MCFS selects  $d$  features base on the coefficients obtained through the regression problem. Another unsupervised feature selection method based on spectral feature selection and sparse learning is described in [6], which shows how to extend the univariate method SPEC [75] to multivariate, making this method capable of removing redundant features.

## 3.2 Wrapper approach

### 3.2.1 Sequential

One of the most outstanding works in this category was introduced by J. G. Dy and C. E. Brodley in [16]. In this work two feature selection criteria were examined: the criterion of maximum likelihood ML (Maximum Likelihood) and the scatter separability criterion (criterion of trace TR). The basic idea of this method is to search through the space of subsets of features, evaluating each candidate subset as follows: Expectation Maximization (EM) [61] or  $k$ -means [60] algorithms are performed on the data described by each candidate subset. Then, the obtained clusters are evaluated with the ML or TR criteria. The method uses a forward selection search for generating subsets of feature that will be evaluated as described above. The algorithm ends when the change in the value of the used criterion is smaller than a given  $\epsilon$ .

Another important work in this category was proposed by Martin H. C. Law et al. in [96]. In this method, the authors assume that features are conditionally independent given the class. The method proposes a strategy to cluster data using the EM algorithm, which was modified to find simultaneously the parameters of the density functions that model the clusters, as well as a set of real values (one for each feature) called “feature saliences” that quantify the relevance of each feature. Features are selected based on these feature saliences values (high feature saliences values are preferred), and the method returns the selected subset of features jointly with the clusters.

In [97] a method where a new optimization criterion that minimizes and maximizes the intra-cluster and inter-cluster inertias, respectively, is proposed. This criterion according to the authors, in most cases is unbiased w.r.t. the number of clusters. The criterion simultaneously provides both a ranking of relevant features and an optimal partition.

Finally, in [98], a method for feature selection called SS-SFS (Simplified Silhouette-Sequential Forward Selection) is proposed. The idea of this method is to select a feature subset that provides the best quality according to the simplified silhouette criteria. In this method, a forward selection search for generating subsets of features is used. Each feature subset is used to describe the data; then the data are clustered using the  $k$ -means clustering algorithm. After the clusters are evaluated by the simplified silhouette criterion, and finally the feature subset that maximizes this criterion is selected.

### 3.2.2 Bio-inspired

Two representative works in this category are those proposed by YongSeog Kim in [99, 100], where an algorithm of evolutionary local selection (ELSA) to search feature subsets as well as the number of clusters using the clustering algorithms  $k$ -means and Gaussian mixture was proposed. Each solution provided by the clustering algorithms is associated with a vector whose elements represent the quality of the evaluation criteria, which are based on the cohesion of the clusters, inter-class separation, and maximum likelihood. Those features of the subset that optimize the objective functions in the evaluation stage are selected as the final result.

Another more recent work, also based on an evolutionary algorithm, is introduced by Dipankar Dutta et al. in [101]. In this work, feature selection is performed while the data are clustered using a multi-objective genetic algorithm (MOGA). The basic idea of this method is to minimize intra-cluster distance (uniformity) and maximize inter-cluster distance (separation) through a multi-objective fitness function. For optimizing this fitness function, the authors employ  $k$ -prototypes [59, 102] as clustering algorithm. Therefore, this method can handle both numerical and non-numerical features (mixed data). In the final stage, this method returns the feature subset that optimizing the fitness function jointly with the clusters.

### 3.3 Hybrid approach

Dash and Liu introduced one of the first unsupervised hybrid feature selection methods in [103]. This method is based on the entropy measure proposed in [69] (filter stage), jointly with the internal scatter separability criterion [16] (wrapper stage). In the filter stage, the authors apply the following search strategy for feature ranking: each feature, in the whole set of features, is removed in turn, and the entropy generated on the data set is calculated. This produces a list of features sorted in according to the degree of disorder that each feature generates when it is removed from the whole set of features. Once all features are sorted, in the wrapper stage, a forward selection search is applied jointly with the  $k$ -means algorithm in order to build clusters which are evaluated by a scatter separability criterion. At the end, the method selects the subset of features that provides the highest value for the separability criterion. According to the authors, this method can be applied to mixed data by performing a discretization of the numerical features before the use of a similarity measure based on the Hamming distance. Finally, It is also worth mentioning that this method performs random sampling of objects, resulting in a loss of information.

Another hybrid method also based on feature ranking was proposed by Yun Li et al. in [104]. In this method, the authors combine an exponential entropy measure with the fuzzy evaluation index FFEI [81] for feature ranking and feature subset selection respectively. The method employs sequential search considering subsets of features based on the generated ranking, by using the fuzzy evaluation index as quality measure. Finally, in the wrapper stage, the fuzzy-c-means algorithm and the scatter separability criterion are employed to select a “compact” subset of features.

Finally, in [105] a hybrid method called BFK which combines  $k$ -means and a Bayesian filter for feature selection is proposed. This method, unlike the above, at the initial stage begins with the wrapper stage, by running the  $k$ -means clustering algorithm on the data set with a range of clusters specified by the user. The clusters are evaluated with the simplified silhouette criterion and that one with the highest value is selected. Subsequently, in the filter stage, to select a subset of features a Bayesian network is built, where each cluster represents a class, the nodes represent features, and the edges represent the relationships between features. Finally, using the concept of Markov blanket a feature subset is selected.

### 3.4 Concluding remarks

From the review on unsupervised feature selection methods reported in the literature, we can see that, there are no studies addressing the problem of unsupervised feature selection specifically for mixed data. And those mentioning that the proposed unsupervised feature selection method can be applied for mixed data, as SUD in [69] and Dash and Liu in [103], have the following limitations:

- a) Although they use different distance measures for each type of feature (Euclidean for numerical features and Hamming for non-numerical ones); for handling mixed data, they recommend performing a preliminary discretization of numerical features. Which, as we have already mentioned, results in loss of information and a high dependence on the used discretization technique [47].
- b) The method proposed by Dash and Liu [103] optimizes an internal validation index, which biases the method towards trivial solutions because it suffers from the so-called “Bias of Criterion Values to Dimension” [16].
- c) In Dash and Liu [103] random sampling of objects is performed, however, according to [7], this type of sampling is not a good choice, because relevant information could be ignored, and additionally the quality of the algorithms may change unpredictably and significantly.

Finally, MOGA [101] due to its meta-heuristic approach based on populations requires the tuning of several parameters (percentage of mutation, crossover rate, the number of iterations, etc.). Moreover, it is also worth mentioning that this algorithm was not proposed exclusively for feature selection, but rather its main objective is clustering, suffering also of the Bias of Criterion Values to Dimension since it also uses an internal validation index for clustering and feature selection.

Table 1: Characteristics of unsupervised feature selection methods for mixed data.

Feature selection method	Bias of Criterion Values to Dimension	Can handle mixed data?	A priori discretization	Selects relevant features	Eliminates redundant features
MOGA (Dutta et al. 2014)	Yes	Yes	No	Yes	No
Dash and Liu, 2000	Yes	No	Yes	Yes	No
SUD (M. Dash et al, 1997)	No	No	Yes	Yes	No
<b>Research proposal</b>	<b>No</b>	<b>Yes</b>	<b>No</b>	<b>Yes</b>	<b>Yes</b>

As we can see, these methods do not solve the unsupervised feature selection problem for mixed data, in addition they have several characteristics that adversely affect the quality of their results. So it is important to propose new methods that pose a solution to these problems. To conclude this section, in Table 1 we can see a comparative of the main differences between the method to be developed in this Ph.D. research and the unsupervised feature selection methods for mixed data mentioned before. In this table, it is possible to appreciate that the method to be developed in this research will work with mixed data without a preprocessing stage, in contrast to SUD and the method proposed by Dash and Liu [103] that perform an a priori feature discretization. Also, unlike other methods, our method will eliminate redundant features and, in the case of using an internal validation index for feature subsets evaluation, we will include a solution to mitigate the Bias of Criterion Values to Dimension.

## 4 Justification and Motivation

As we have previously stated, there are several real-world applications where data are described by both, numerical and non-numerical features, that is, mixed data; and in many cases, these data are not labeled (unsupervised data). However, as we have seen in the related work section, the development of unsupervised feature selection methods for mixed data has been little explored. To the best of our knowledge, in the literature, only three works raise the possibility of processing mixed data. However, as we have already discussed, two of these methods perform an a priori discretization of numerical features. Meanwhile, the third work is not really a feature selection method, because its primary objective is data clustering.

In view of the above mentioned, in this Ph.D. research, we will introduce a new unsupervised feature selection method for mixed data.

## 5 Ph.D. Research Proposal

In this section, we address the problem to be solved, the research questions, hypothesis, objectives, the expected contributions, the methodology, the publication plan and the schedule of activities.

### 5.1 Problem to be solved

The problem to be solved in this Ph.D. research is to develop a new unsupervised feature selection method for mixed data, which outperforms the methods reported in the literature. This raises some serious difficulties, which are listed below:

1. In unsupervised feature selection, there is not a measure to assess the relevance of a feature (or a feature subset) in mixed datasets. Therefore new mechanisms for identifying relevant features in this kind of data should be introduced.
2. For identifying redundant features, a pairwise comparison between features is commonly performed. However, if the features are of different type, it is not clear how to determine redundant features.
3. Simultaneously identifying relevant and non-redundant feature subsets in unsupervised mixed data has not been studied.

### 5.2 Research questions

From the difficulties above commented, the following research questions emerge:

- Q1.** How to quantify the relevance of a feature (or a feature subset) in unsupervised mixed datasets?
- Q2.** How to identify redundant features in unsupervised mixed datasets?
- Q3.** How to select relevant and non-redundant features in unsupervised mixed datasets?

### 5.3 Hypothesis

In our research, we consider the following hypothesis:

*It is possible to find a subset of relevant and non-redundant features in unsupervised mixed datasets.*

## 5.4 Objectives

Following our hypothesis, to solve the research questions we established the following objectives.

### 5.4.1 Main objective

Propose a new unsupervised feature selection method that can select a subset of relevant and non-redundant features in mixed data, which outcomes statistically better results than state-of-the-art methods.

### 5.4.2 Specific objectives

- Propose a quality measure to evaluate how relevant a feature (or a feature subset) is in unsupervised mixed datasets.
- Propose a measure to evaluate redundancy between features in unsupervised mixed datasets.
- Propose a way for combining redundancy and relevance measures for selecting features in unsupervised mixed datasets.
- Develop a new unsupervised feature selection method for mixed data such that the results, in terms of classification, be statistically better than the state-of-the-art methods.

## 5.5 Methodology

To carry out our specific objectives, the following methodology is proposed:

- I. Select, collect and critically analyze the related works reported in the literature.
- II. Data generation and collecting.
  1. Collect and pre-process real-world datasets from different sources to evaluate the performance of the unsupervised feature selection method developed in this research.
  2. Generate synthetic datasets with different characteristics to evaluate the performance of the proposed method in controlled situations. These datasets will be generated following the most commonly used guidelines in feature selection [16, 96, 106] and clustering [107].
- III. Develop a strategy for identifying and selecting relevant features in unsupervised mixed datasets, for doing this, we will follow the next steps:
  1. First, we will analyze and evaluate the possibility of using or extending the ideas employed by the reported supervised feature selection methods for mixed data in an unsupervised context. At this point, we will explore among others: Information theory [43, 47, 108–110], Testor’s theory [111, 112], Rough set theory [113–118], and the decision tree approach [119].
  2. Develop a way to assess feature subsets and/or individual features based on the most effective and important approaches used for identifying relevant features in unsupervised feature selection. At this point, we will address, among others, the following lines of research:

- a) **Spectral feature selection.** In this approach, we consider proposing new algorithms to construct a good data representation through a kernel matrix [120] (which is necessary for constructing Laplacian matrices). In this step, we will evaluate the performance of some kernel functions [42, 121] and distance measures, such as those described in [122–125], that have been developed for mixed data. Also, we will explore new score functions to measure the relevance of a feature following this approach.
  - b) **Decision trees.** Here, we will propose a way for evaluating features individually using split evaluation measures and strategies used in unsupervised decision trees [126, 127] for identifying relevant features in mixed data.
  - c) **Information theory.** In this approach, initially, we will propose a way for evaluating feature relevance in unsupervised mixed datasets using measures from information theory [68] such as entropy, mutual information, or conditional mutual information.
3. Propose a new method for identifying and selecting relevant features in unsupervised mixed data based on the experience obtained from points III1 and III2.
  4. Implementing the main unsupervised feature selection methods of the approaches mentioned above for identifying relevant features.
  5. Testing, comparing results, analysis, and feedback of the approaches and methods analyzed in the points III1, III2, III3, and III4 as well as the new ideas generated from these steps.
- IV. Develop a strategy for identifying redundant features in unsupervised mixed datasets, for this, the next steps will be followed:
1. Analyze and evaluate the possibility of extending, to the unsupervised context, the strategies used in supervised feature selection for identifying redundant features in mixed data. For doing this, we will explore some strategies used in Game Theory [128, 129] and Information Theory [10, 11].
  2. Develop a way to assess feature redundancy in mixed datasets following ideas from the most effective and robust multivariate strategies used for eliminating redundant features in unsupervised numerical datasets. At this point, we will address three main approaches:
    - a) **Multivariate spectral feature selection.** It is planned to develop a mechanism for identifying redundant features using the kernel matrices constructed in point III(2)a applying the spectral feature selection approach in a multivariate way [6].
    - b) **Sparse learning approach.** Develop an evaluation mechanism based on the sparse learning [87, 93] approach to identify redundant features in mixed data, modeling the feature selection as a loss minimization problem using the similarity matrix constructed in point III(2)a.
    - c) **Statistical approach.** Evaluate non-parametric and functional dependence methods [7] for measuring the similarity (correlation) between features in mixed data.
  3. Based on the experience obtained from points IV1 and IV2, proposing a new algorithm for identifying and eliminating redundant features in unsupervised mixed data.
  4. Implementing the main unsupervised feature selection methods of the approaches mentioned above for eliminating redundant features.
  5. Testing, comparing results, analysis, and feedback of the approaches and methods analyzed in the points IV1, IV2, IV3, and IV4 as well as the new ideas generated in these steps.

- V. Proposing a new unsupervised feature selection method for identifying and selecting relevant and non-redundant features in mixed datasets. For this, we will proceed as follows:
1. Propose a strategy to combine the results obtained at points [III](#) and [IV](#) for selecting a feature subset. In this step, filter, wrapper, and hybrids approaches will be explored considering, among others, the following combination options:
    - a) **Sequential combination.** Propose an algorithm for generating and evaluating feature subsets using the measures of relevance and redundancy developed in the points [III](#) and [IV](#), considering some sequential or bio-inspired search strategies.
    - b) **Joint combination.** Propose an algorithm that allows the interaction between [III](#) and [IV](#) for a joint search of relevant and non-redundant features. Additionally, we will look for a function that simultaneously quantifies the relevance/redundancy of the features in mixed data.
  2. Testing, comparison of results, analysis, and feedback.

VI. Experimental evaluation.

1. Develop a platform for feature selection, which allows doing the evaluation, and comparative study of unsupervised feature selection methods.
2. Evaluate and compare the proposed unsupervised feature selection method against state-of-the-art methods. For this evaluation and comparison, we will follow two standard forms of assessment of unsupervised feature selection methods:
  - a) In terms of the clustering quality [[27](#), [90](#), [94](#), [130–132](#)]. In this context, we will use the most common evaluation measures in unsupervised classification such as ACC, NMI or the Jaccard coefficient, over the results of a clustering algorithm like  $k$ -prototypes or finite mixture models; working over the selected subset of features.
  - b) In terms of classification quality using supervised learning algorithms [[54](#), [75](#), [84](#), [85](#), [87](#)], such as Support Vector Machines, Random-Forest,  $k$  nearest-neighbor, Naive Bayes, and Decision Trees; working over the selected subset of features.

A diagram of this methodology is shown in [Figure 7](#).

## 5.6 Expected contributions

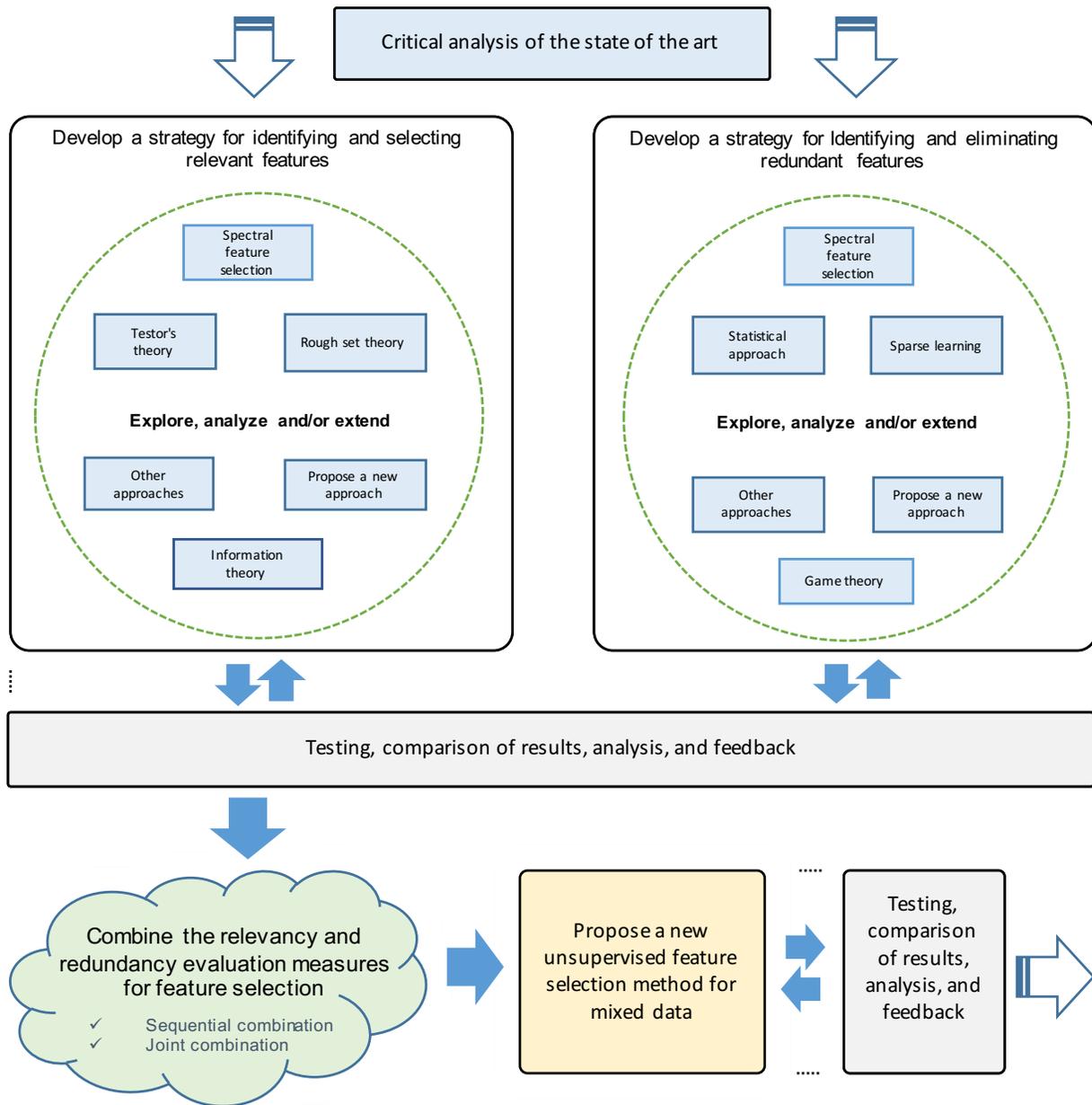
The expected contributions of this research include:

- A new unsupervised feature selection method for mixed data.
- A new measure to assess feature relevance in unsupervised mixed data.
- A new measure to quantify the redundancy between features in unsupervised mixed data.

## 5.7 Publication plan

The publication plan of this research is as follows:

1. A review of unsupervised feature selection methods. This will be sent to [Artificial Intelligence review](#).



Combine the relevancy and redundancy evaluation measures for feature selection

- ✓ Sequential combination
- ✓ Joint combination

➔

Propose a new unsupervised feature selection method for mixed data

↔

Testing, comparison of results, analysis, and feedback

➔

Figure 7: General diagram of the proposed methodology.

2. An article in a JCR Journal reporting how to identify and select relevant features in unsupervised mixed datasets. This will be sent to *Pattern Recognition* or *Expert Systems with Applications*.
3. An article in a JCR Journal reporting a way to identify and eliminate redundant features in unsupervised mixed datasets. This article will be sent to *Pattern Analysis and Applications* or *Knowledge and Information Systems*.
4. An article in a JCR Journal reporting a new unsupervised feature selection method for mixed data

that eliminates redundant features and selects the most relevant ones. This article may be sent to *NeuroComputing*, or *Knowledge-Based Systems*.

- We also plan to send articles to refereed international conferences (for example: [MCPR](#), [CIARP](#), and/or [MICA](#)) reporting intermediate results.

## 5.8 Schedule

Table 2 shows the schedule for this Ph.D. research.

Table 2: Schedule of activities.

No.	Activity	Quarter	2016			2017			2018			2019			
			1	2	3	1	2	3	1	2	3	1	2	3	
1	Collection and analysis of the literature (Point I of the methodology)			✓											
2	Collection and generation of mixed datasets			✓											
3	Developing preliminary results			✓											
4	Writing the PhD. research proposal			✓											
5	Public defense of the PhD. research proposal														
6	Critical study, analysis and exploration of the most effective and important approaches used for identifying relevant features (Point III of the methodology).			✓											
7	Develop a strategy for identifying and selecting relevant features in mixed data sets			✓											
8	Critical study and analysis of feature selection approaches for identifying redundant features (Point IV of the methodology)				✓										
9	Develop a strategy to identifying and eliminating redundant features in mixed data														
10	Propose an unsupervised feature selection method for mixed data that eliminates redundant features and selects the relevant ones (Point V of the methodology)														
11	Experimental evaluation (Point VI of the methodology)			✓											
12	Writing of the PhD. thesis document				✓										
13	Writing and submitting articles to journals or conferences				✓										
14	Revisions of the PhD. thesis document (advisors)														
15	Delivery of the PhD. thesis document to the PhD. committee														
16	Thesis corrections														
17	Public defense of the results of this PhD. research														🎓

Period of time 

In progress 

PhD. defense 

## 6 Preliminary Results

This section presents the preliminary results obtained following the methodology described in subsection 5.5. Section 6.1 presents a new filter unsupervised feature selection method for mixed data, and Section 6.2, we present our experimental results over real and synthetic datasets.

### 6.1 A new filter unsupervised spectral feature selection method for mixed data

Given a collection of  $m$  objects  $X_T = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ , described by a set of  $n$  features  $T = \{F_1, F_2, \dots, F_n\}$ , and a  $m \times m$  similarity matrix  $W$ , containing the similarities  $w_{ij} \geq 0$  between all pairs of objects  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . According to the spectral clustering theory [74, 133, 134], we can obtain structural information (cluster structure) from  $X_T$ , by studying the eigensystem of the Laplacian matrix derived from  $W$ . Based on this knowledge, spectral feature selection [6] is a relatively new and effective feature selection framework that uses the similarity matrix  $W$ , the Laplacian matrix, and the eigen-system of the Laplacian matrix for finding relevant features through the features' consistency. Nevertheless, to quantify the feature relevance in mixed data using the theoretical bases of this feature selection framework, two important problems must be solved: 1) a way of quantifying the objects' similarities in mixed data must be proposed. 2) in mixed data, there is no way to measure the consistency of a non-numerical feature against the Laplacian matrix or its eigenvectors (as it is usually done). To address these problems we propose the following:

1. Build a similarity matrix  $W$  through a kernel function capable of quantifying the similarity between objects in mixed datasets. With this, in addition to employing an effective and widely used approach in the context of unsupervised learning for describing relationships among objects, we also provide a mechanism for abstracting the feature type in mixed data.
2. Because the structural information of the data is contained in the spectrum of the Laplacian matrix, and since the spectrum is independent of the type of features in the dataset (because the spectrum is in function of the  $W$  matrix). We propose measuring the consistency (relevance) of each feature through a feature relevance evaluation measure based on the spectrum of the Normalized Laplacian matrix  $\mathcal{L}$  derived from the similarity matrix  $W$  together with a leave-one-out feature elimination strategy. The idea is to quantify the variation of the distribution of the spectrum in such a way that depending on that variation; we can measure the consistency of each feature  $F_i \in T$ .

Below, these points are explained in detail.

#### 6.1.1 The similarity matrix

A  $m \times m$  object similarity matrix  $W = w(i, j) = w_{ij}$ , contains the similarity between the objects in a dataset  $X_T$ . In this matrix, the elements  $w_{ij}$  represent the similarity between the object  $\mathbf{x}_i$  and the object  $\mathbf{x}_j$ , which can be viewed as the weight on the edge connecting the  $i^{th}$  and  $j^{th}$  data points in a graph  $G$ . The number of edges on  $G$  depends on the type of the graph that we want to model, such as the  $k$ -nearest neighbor or the fully connected graph [133]. To model the global structure of the data, we construct the fully connected graph, and we take  $W$  as the adjacency matrix of  $G$ .

$W$  is also called as affinity or kernel matrix, a positive semi-definite matrix where it holds that  $\forall \mathbf{z} \in \mathbb{R}^m$ ,  $\mathbf{z}^T W \mathbf{z} \geq 0$ . For building the matrix  $W$ , we propose to use the clinical kernel [121], which assumes that all features are equally important in the original space. We selected this kernel because it has shown good

results for classifying mixed data [42, 135]. The clinical kernel averages univariate sub-kernels for each feature as follows:

$$w_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n} \sum_{p=1}^n k(x_{ip}, x_{jp}) \quad (9)$$

where  $k(a, b) = 1 - \text{dist}_F(a, b)$ , being  $\text{dist}_F(a, b)$  any metric that can handle different types of features.

### 6.1.2 The Normalized Laplacian matrix and its spectrum

Let  $\mathbf{d}$  denote the vector:  $\mathbf{d} = \{d_1, d_2, \dots, d_m\}$ , where  $d_i = \sum_{k=1}^m w_{ik}$ . The degree matrix  $D$  of the graph  $G$  is defined as  $D(i, j) = d_i$  if  $i = j$ , and 0 otherwise.  $d_i$  can be interpreted as an estimation of the density around  $\mathbf{x}_i$ . The Laplacian ( $L$ ) and the Normalized Laplacian ( $\mathcal{L}$ ) matrices are defined as:

$$\begin{aligned} L &= D - W \\ \mathcal{L} &= D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \end{aligned} \quad (10)$$

where  $D^{-\frac{1}{2}}$  is the reciprocal square root of  $D$  whose diagonal entries are the reciprocals of the positive square roots of the diagonal entries of  $D$ . Finally, the spectrum  $S_{\mathcal{L}}$  of the Normalized Laplacian matrix  $\mathcal{L}$  is defined as:

$$S_{\mathcal{L}} = (\lambda_1, \lambda_2, \dots, \lambda_m) \quad (11)$$

where  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$  are the eigenvalues of  $\mathcal{L}$  arranged in ascending order. In this spectrum, the eigenvalue  $\lambda_1$  is always equal to 0 [74], and together with its corresponding eigenvector, they are called the trivial eigenpair of  $\mathcal{L}$ .

$S_{\mathcal{L}}$  has two useful properties for feature selection. 1) the eigenvalues close to zero (i.e., the first non-trivial eigenvalues of  $S_{\mathcal{L}}$ ) quantify the consistency of their corresponding eigenvectors; in other words, they quantify how well their corresponding eigenvectors can separate the data [6]. 2) when a dataset has a separable cluster structure, the differences between elements in  $S_{\mathcal{L}}$  (also called eigengaps or spectral gaps) are larger [133]. For illustrating these facts, in Figure 8 we show the distribution of the spectrum  $S_{\mathcal{L}}$  of a synthetic<sup>3</sup> mixed dataset composed by 600 objects, 44 relevant features, and six clusters. In Figure 8a, we can see the spectrum of this dataset when it is described only by the 44 relevant features. While in Figure 8b, we can observe the same dataset, but adding 20 irrelevant features. As can be seen in these figures, when we added irrelevant features, the spectral gap between the first non-trivial eigenvalues of  $S_{\mathcal{L}}$  tend to be smaller.

The two properties above mentioned can help us to identify relevant features in the data. Therefore, we propose to use these properties for introducing a new unsupervised feature selection method capable of handling mixed data.

### 6.1.3 Identifying relevant features

As we have already commented, the cluster structure of the data can be quantified from the information provided by the first eigenvalues<sup>4</sup> of  $S_{\mathcal{L}}$ . Based on this fact, and also knowing that the first  $k+1$  eigenvectors

<sup>3</sup>For details on how these datasets are generated, see Section 6.2.

<sup>4</sup>Except the eigenvalue  $\lambda_1$ , which, as we have already mentioned, is always equal to 0.

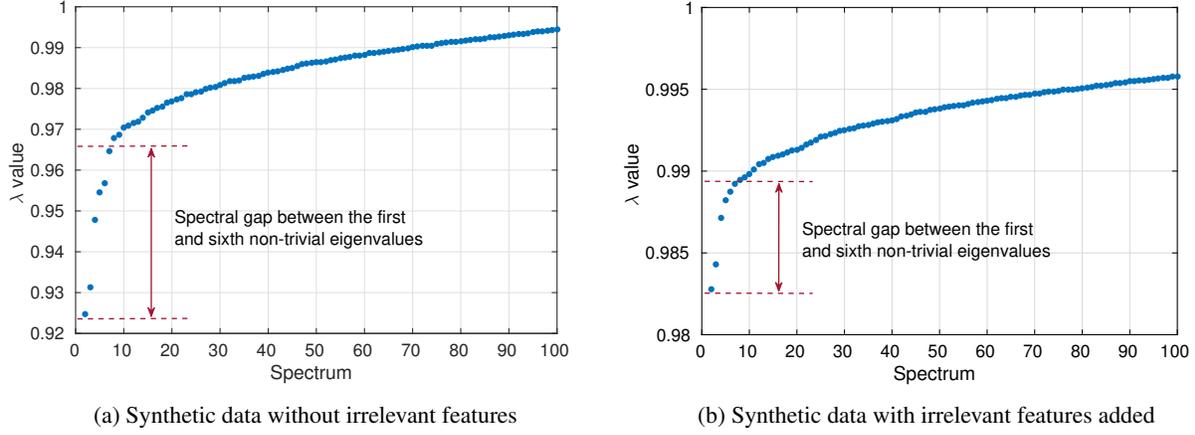


Figure 8: Distribution of the spectrum  $S_{\mathcal{L}}$  (100 non-trivial eigenvalues) of a synthetic mixed dataset.

of  $\mathcal{L}$  describe an optimal partition on  $X_T$ , separating the data in  $k$  clusters [6]. We define the spectral gap score for  $X_T$ , being  $k$  the number of clusters in  $X_T$ , as follows:

$$\gamma(X_T, k) = \sum_{i=2}^k \sum_{j=i+1}^{k+1} \left| \frac{\lambda_i - \lambda_j}{\tau} \right|; \quad (12)$$

where  $\tau = \sum_{i=2}^{k+1} \lambda_i$  is a normalization term.  $\lambda_i, i = 2, \dots, k+1$  are the first  $k$  nontrivial eigenvalues of the spectrum  $S_{\mathcal{L}}$ . Using this score, the bigger the gap of the first  $k+1$  eigenvalues of  $S_{\mathcal{L}}$ , the best is the cluster structure defined in the data.

Based on (12), we propose quantifying the relevance of each feature  $F_i \in T$ , as follows:

$$\varphi(F_i) = \gamma(X_T, k) - \gamma(X_{T_i}, k) \quad (13)$$

where  $X_{T_i}$  denotes the dataset described by the set of features  $T_i$ , which contains all features except  $F_i$ . The idea is to measure the contribution of the feature  $F_i$  in the definition of the cluster structure that might exist in  $X_T$ . As we can see, this feature evaluation measure results in two possible cases:

- **Case 1.**  $\varphi(F_i) > 0$ . In this case,  $F_i$  is a relevant feature, since when it is eliminated the spectral gap decreases.
- **Case 2.**  $\varphi(F_i) \leq 0$ . In this case,  $F_i$  is an irrelevant feature, since when  $F_i$  is removed from the dataset the spectral gap is greater or equal than using all features.

Notice that  $\varphi(\cdot)$  does not depend on the feature type. Therefore this feature relevance evaluation measure can be applied over numeric and non-numeric features.

Our proposed method begins with the construction of the similarity matrix  $W$  from the original dataset  $X_T$ . From  $W$ , we construct the Normalized Laplacian matrix  $\mathcal{L}$  along with its spectrum  $S_{\mathcal{L}}$ . Later, using (12), the spectral gap score  $\gamma(X_T, k)$  is computed. This procedure is repeated  $n$  times using a leave-one-out feature elimination strategy over  $T$  to quantify the relevance of each feature  $F_i$  through (13). Finally, the features in  $T$  are ranked from the most to the least relevant according to the values obtained by  $\varphi(\cdot)$ . The

<p><b>Input</b> : <math>X : m \times n</math> dataset, with <math>m</math> objects and <math>n</math> features  <math>k</math>: the desired number of clusters</p> <p><b>Output</b>: <math>F_{Rank}, w</math>: Feature ranking and feature weights</p> <ol style="list-style-type: none"> <li>1 <math>T = \{F_1, F_2, \dots, F_n\}</math>;</li> <li>2 Build <math>W</math>, the similarity matrix from <math>X_T</math>;</li> <li>3 Build <math>\mathcal{L}</math> from <math>W</math>;</li> <li>4 Calculate the spectrum of <math>\mathcal{L}</math> and get <math>S_{\mathcal{L}}</math>;</li> <li>5 <math>\gamma_T \leftarrow \gamma(X_T, k)</math>;</li> <li>6 <b>for</b> <math>i \leftarrow 1</math> <b>to</b> <math>n</math> <b>do</b></li> <li style="padding-left: 20px;">7 <math>T_i \leftarrow T \setminus F_i</math>;</li> <li style="padding-left: 20px;">8 Build <math>W</math>, the similarity matrix from <math>X_{T_i}</math>;</li> <li style="padding-left: 20px;">9 Build <math>\mathcal{L}</math> from <math>W</math>;</li> <li style="padding-left: 20px;">10 Calculate the spectrum of <math>\mathcal{L}</math> and get <math>S_{\mathcal{L}}</math>;</li> <li style="padding-left: 20px;">11 <math>\gamma_{T_i} \leftarrow \gamma(X_{T_i}, k)</math>;</li> <li style="padding-left: 20px;">12 <math>w[i] \leftarrow \varphi(F_i) = \gamma_T - \gamma_{T_i}</math>;</li> <li>13 <b>end</b></li> <li>14 Sort <math>w</math> in descending order and build <math>F_{Rank}</math> according this order;</li> </ol>
--

**Algorithm 1:** Unsupervised Spectral Feature Selection method for Mixed data (USFSM).

pseudo-code of the proposed filter method, which we have called USFSM (Unsupervised Spectral Feature Selection method for Mixed data), is shown in Algorithm 1.

### 6.1.4 Time complexity analysis

For analyzing the time complexity of USFSM for an  $X_T$  with  $m$  objects,  $n$  features, and  $k$  clusters, we take into account the following. In lines 2-5 as well as in lines 8-11 of the Algorithm 1 we need to perform the next steps: Build  $W$  from  $X_T$  that is  $O(nm^2)$ . Build  $\mathcal{L}$  from  $W$ , which is  $O(m^2)$ . Calculate the spectrum of  $\mathcal{L}$ , which is  $O(m^3)$ . And finally, calculate  $\gamma(\cdot, \cdot)$  that is  $O(k^2)$ . Then the time complexity of this part (lines 2-5 or lines 8-11) is  $T_{2-5} = O(nm^2) + O(m^2) + O(m^3) + O(k^2) = O(m^3)$ .

Lines 8-11, which complexity is  $T_{2-5}$  are repeated  $n$  times, therefore we have a time complexity of  $T_{6-12} = nT_{2-5} = O(nm^3)$ . The step 14 sorts  $n$  features which is  $T_{14} = O(n \log(n))$  (using merge sort). Therefore, the total time complexity for our method is:

$$T_{USFSM} = T_{2-5} + T_{6-12} + T_{14} = O(m^3) + O(nm^3) + O(n \log(n)) = O(nm^3)$$

## 6.2 Experimental results

To evaluate the performance of USFSM, two types of experiments were carried out. In the first experiment, the performance of the proposed method was evaluated in term of clustering results using two clustering algorithms for mixed data:  $k$ -prototypes and Finite Mixed Models using EM as optimization algorithm. The evaluation was performed according to the steps and validation measures (ACC and NMI) described in Section 2.3. For ACC evaluation measure, the best match with the true labels was found using the Kuhn-Munkres<sup>5</sup> algorithm [136]. In the second experiment, we evaluate our method in terms of

<sup>5</sup><http://www.cad.zju.edu.cn/home/dengcai/Data/code/hungarian.m>

supervised classification using four well-known classifiers: SVM [137], KNN [138] ( $k = 3$ ), Naive Bayes (NB) [139], and Random-Forest [140], following the evaluation strategy described in Section 2.3 and applying stratified 10-fold cross-validation. These four classifiers were chosen because they represent four quite different approaches for supervised classification, and they are among the most used classifiers for validation of unsupervised feature selection methods. The clustering algorithms and classifiers above mentioned were taken from the Weka data mining software package [141], using the default parameter values.

Table 3: Details of the real mixed datasets from the UCI repository used in our experiments.

#	Dataset	Number of objects	Features			Number of classes
			Numerical	Non-numerical	All	
1	Acute-Inflammations	120	1	5	6	2
2	Automobile	205	15	11	26	7
3	Contraception	1473	2	8	10	3
4	Flags	194	10	20	30	8
5	Heart-c	303	6	7	13	2
6	Heart-h	294	6	7	13	2
7	Horse-colic	368	7	16	23	2
8	Post-operative	90	1	7	8	3
9	Teaching-Assist-Eval	151	1	4	5	3
10	Thoracic-Surgery	470	3	14	17	2
11	Bridges-version 2	105	2	10	12	6
12	Credit-approval	690	6	10	16	2
13	Credit-German	1000	7	14	21	2
14	Cylinder-bands	540	18	22	40	2
15	Dermatology	366	1	34	35	6
16	Heart-statlog	270	6	7	13	2
17	Hepatitis	155	7	13	20	2
18	Labor	57	8	8	16	2
19	Liver-disorders	345	6	1	7	2
20	Tae	151	3	3	6	3

In both experiments, 20 real datasets taken from the UCI repository [142], which have been extensively used for validation of supervised and unsupervised feature selection methods were employed. The characteristics of these datasets are shown in Table 3. Also, we have generated 15 synthetic mixed datasets following the guidelines described in [16, 96] for the generation of synthetic numerical features; and [106] for the generation of synthetic non-numerical features. Relevant numerical features were generated following a multivariate normal distribution defined for each cluster, while irrelevant features were generated following a uniform distribution through all clusters. Likewise, relevant non-numerical features follow a multinomial distribution defined for each cluster; meanwhile, irrelevant ones have exactly the same multinomial distribution in all the clusters. Then, each synthetic dataset is a mixture of Gaussians-Multinomials. The parameter values used to generate these datasets appear in Table 4. In this table, we can see that the number of features, clusters, and objects for each synthetic dataset were set in different ranges. Also, this table includes the parameter values used for generating relevant features (means, covariances and the number of values). These values allow us evaluating the ability of the feature selection methods for selecting features under different degrees of overlapping between clusters.

Since the method developed so far in this Ph.D. research proposal is a univariate filter method based

on ranking, we perform the comparison against filter methods of this same type and filter methods where the number of features to select can be specified in advance. Therefore, in our experiments, we made this comparison against the following nine state-of-the-art unsupervised filter feature selection methods: SUD [69], Laplacian score<sup>6</sup> [67], SVD-Entropy<sup>7</sup> [71] (using simple ranking), SPEC<sup>8</sup> [75], UFSACO [84], MGSACO [85], IRRFSACO, RRFSAO<sup>9</sup> [54], and FSFS [79]. In all cases except SUD, we used the author’s implementation of these methods with the parameters recommended by their respective authors. Meanwhile, SUD was implemented based on to the description provided by their authors in [69], and using simple binning [141] as discretization method.

For our method, we use the feature comparison function used in the Heterogeneous Euclidean-Overlap Metric (HEOM) [124], which is defined as follows:

$$dist_F(a, b) = \begin{cases} 1 & \text{if } a \text{ or } b \text{ are unknown} \\ overlap(a, b) & \text{if } F \text{ is non-numerical} \\ diff_F(a, b) & \text{if } F \text{ is numerical} \end{cases} \quad (14)$$

where  $overlap(x, y)$  is 0 if  $x = y$  and 1 otherwise; and  $diff_F(x, y) = \frac{|x-y|}{max_F - min_F}$ , being  $max_F$  and  $min_F$  the maximum and minimum values of the feature  $F$  in the dataset, respectively.

In our experiments, all datasets were standardized, and the missing values for non-numerical, and numerical features were replaced by the modes and means respectively. For all the datasets, class labels were removed for feature selection and clustering. Furthermore, as usual, for the feature selection methods that can only process numerical features the non-numerical features were transformed to numerical ones by mapping each categorical value into an integer value in the order of appearance of the dataset.

---

<sup>6</sup><http://www.cad.zju.edu.cn/home/dengcai/Data/code/LaplacianScore.m>

<sup>7</sup><http://horn.tau.ac.il/compact.html>

<sup>8</sup><http://featureselection.asu.edu/old/software.php>

<sup>9</sup><http://kfst.uok.ac.ir/software.html>

Table 4: Mixed Gaussians-Multinomials synthetic datasets.

Dataset	Features				No. of Objects	No. of clusters	Objects distribution for cluster	Rel. Num. Feature parameters	Rel. Nom. Feature parameters	
	Non-numerical		Numerical							All
	Rel	Irrel	Rel	Irrel						
$S_1$	8	12	2	3	25	400	2	C1=C2=200		
$S_2$	10	14	2	4	30	502	2	C1=210, C2=292	- Means $\mu$ , were sampled from a uniform distribution on $[-5, 5]$ .	
$S_3$	16	23	3	23	65	498	3	C1=C2=C3=166	- The number of values for each feature $F_i$ was sampled from a uniform distribution on $[2, 5]$ .	
$S_4$	16	23	3	23	65	548	4	C1=110, C2=191, C3=137, C4=110	-The elements of the diagonal covariance matrices $\sigma$ were sampled from a uniform distribution on $[0.7, 4]$ .	
$S_5$	18	27	2	28	75	550	5	C1=110, C2=77, C3=110, C4=176, C5=77	- Probability distribution of $F_i$ for each cluster was sampled from uniform distribution on $[0.0, 1.0]$ .	
-----										
$S_6$	24	36	2	38	100	658	2	C1=383, C2=275		
$S_7$	36	54	3	57	150	656	4	C1=191, C2=164 C3=164, C4=137	- Means $\mu$ were sampled from a uniform distribution on $[-5, 5]$ .	
$S_8$	41	61	3	65	170	600	6	C1=138, C4=100, C5=119, C2=C3=C6=81	- The number of values for each feature $F_i$ was sampled from a uniform distribution on $[2, 8]$ .	
$S_9$	48	71	4	76	200	595	7	C1=C4=101, C2=C7=69, C3=C5=C6=85	-The elements of the diagonal covariance matrices $\sigma$ were sampled from a uniform distribution on $[0.7, 8.0]$ .	
$S_{10}$	56	82	4	88	230	600	4	C1=179, C2=121, C3=C4=150	- Probability distribution of $F_i$ for each cluster was sampled from uniform distribution on $[0.0, 1.0]$ .	
-----										
$S_{11}$	30	45	4	71	150	600	4	C1=240, C3=150, C2=C4=105		
$S_{12}$	17	25	2	16	60	400	2	C1=219, C2=181 C1=C5=C6=C9=60,	- Means $\mu$ were sampled from a uniform distribution on $[-5, 5]$ .	
$S_{13}$	34	51	5	80	170	600	10	C2=C4=C7=C10=42, C3=114, C8=78	-The elements of the diagonal covariance matrices $\sigma$ were sampled from a uniform distribution on $[0.7, 8.0]$ .	
$S_{14}$	42	61	6	99	210	696	6	C1=C2=150, C3=C6=88, C4=C5=116	- Probability distribution of $F_i$ for each cluster was sampled from uniform distribution on $[0.0, 1.0]$ .	
$S_{15}$	46	69	5	100	230	700	2	C1=455, C2=245		

In order to perform the comparison of the results obtained by the proposed method against the results produced by the other tested methods over the used datasets, we performed a two-sided Wilcoxon rank sum test [143] using a confidence level of 95%. We use this statistical test because it is one of the most used to evaluate classification and clustering results. By applying the Wilcoxon test, we detect if the proposed method gets a result with a statistically significant difference respect to the other methods. Symbols “+” and “-” indicate statistically significant better or worse behavior respectively. In the tables showing our results, the best method on average for each dataset appears in “**bold**”, and the last row of these tables show the average of the evaluation measure over all tested datasets.

All experiments were run in Matlab<sup>®</sup> R2016a, using a computer with an Intel Core i7-5820k 3.30 GHz processor with 32 GB DDR4 RAM, running 64-bit GNU/Linux. The implementation of our method was done in Java 1.8.0\_92, using the Apache Commons Math library for matrix operations and the eigen-system computation.

### 6.2.1 Evaluation in terms of clustering

In this experiment, we evaluate our unsupervised feature selection method in terms of clustering results; for this propose, we use the first 20% of the ranked<sup>10</sup> features for the methods based on ranking (USFSM, SUD, Laplacian score, SVD-Entropy, SPEC). Meanwhile, for the other methods (UFSACO, MGSACO, IRRFSACO, RRFSAO and FSFS), we fix the number of features to select as 20% of the whole set of features. The clustering results are evaluated through the ACC and NMI measures following the steps described in Section 2.3. For each tested dataset, the  $k$ -prototypes as well as the EM clustering algorithms were repeated 10 times with random initializations, and we fixed the number of clusters as the number of classes of each dataset.

Table 5: NMI results of  $k$ -prototypes on 20 mixed datasets from the UCI repository.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
Acute-Inflamations	0.025+	0.018+	0.141+	0.044+	0.044+	0.044+	0.044+	0.190+	0.414	<b>0.447</b>
Automobile	0.183+	0.159+	0.165+	0.159+	0.181+	0.162+	0.145+	0.225+	0.189+	<b>0.264</b>
Contraception	0.022	0.015+	0.015+	<b>0.031-</b>	0.025	<b>0.031-</b>	<b>0.031-</b>	0.021	0.022	0.022
Flags	0.178-	0.134+	0.171-	0.146	<b>0.277-</b>	0.175	0.132+	0.178	0.125+	0.158
Heart-c	0.125+	<b>0.299-</b>	0.077+	0.116+	0.156+	0.121+	0.120+	0.115+	0.195+	0.262
Heart-h	0.256	0.013+	0.046+	0.038+	0.083+	0.086+	0.012+	0.118+	0.203+	<b>0.285</b>
Horse-colic	0.050-	0.03	0.020+	0.027+	0.021	0.029	0.010+	0.053	<b>0.051-</b>	0.032
Post-Operative	0.022	0.017	0.018	0.011	<b>0.025-</b>	0.024-	0.023-	0.022	0.018	0.016
Teaching-Assist-Eval	0.051	0.016+	<b>0.048</b>	0.047	0.044	0.047	0.047	0.039	0.019+	0.045
Thoracic Surgery	<b>0.006-</b>	0.002+	0.003+	0.001+	0.003+	0.001+	0.001+	0.004	0.001+	0.004
Bridges-version-2	0.334+	0.286+	0.381	0.215+	0.115+	0.109+	0.096+	0.262+	0.307+	<b>0.378</b>
Credit-approval	0.113+	0.000+	0.023+	0.005+	0.016+	0.025+	0.012+	<b>0.295</b>	0.070+	0.276
Credit-german	0.010-	0.001	0.005-	0.006	0.008-	0.006	<b>0.014-</b>	0.006-	0.007-	0.002
Cylinder-bands	<b>0.019-</b>	0.013-	0.016-	0.005	0.004+	0.005	0.001+	0.008	0.011	0.008
Dermatology	0.474+	0.375+	0.368+	0.377+	0.471	0.492+	0.478+	0.362+	0.288+	<b>0.547</b>
Heart-statlog	0.116+	0.303-	0.028+	0.075+	0.094+	0.066+	0.077+	0.151	0.18	<b>0.202</b>
Hepatitis	<b>0.146-</b>	0.086	0.075	0.052+	0.053+	0.049+	0.034+	0.051+	0.087	0.103
Labor	0.153	0.135	<b>0.184</b>	0.093	0.096	0.078+	0.094	0.051+	0.122	0.155
Liver-disorders	0.002	<b>0.010-</b>	0.007-	0.005-	0.003	0.005-	0.005-	0.008-	0.002	0.002
Tae	0.051	0.016+	0.047	0.047	0.031	0.047	0.047	0.039	0.015+	<b>0.051</b>
<b>Average</b>	<b>0.117</b>	<b>0.096</b>	<b>0.092</b>	<b>0.075</b>	<b>0.087</b>	<b>0.080</b>	<b>0.071</b>	<b>0.110</b>	<b>0.116</b>	<b>0.163</b>

Tables 5 and 6 show the results of the comparison between the proposed USFSM method and the other unsupervised filter methods in terms of NMI and clustering accuracy (ACC), using  $k$ -prototypes over the real datasets of Table 3, respectively. From these tables, we can see that the proposed method outperforms on

<sup>10</sup>We have set this percent value to assess the ability each feature selection method has to place the most relevant features at the beginning of the ranking. Other percentages were also considered, see experiments in this same subsection.

Table 6: Clustering Accuracy (ACC) results of  $k$ -prototypes on 20 mixed datasets from the UCI repository.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
Acute-Inflamations	0.583+	0.575+	0.655+	0.620+	0.620+	0.620+	0.620+	0.715	0.826	<b>0.829</b>
Automobile	0.402	0.367+	0.376+	0.366+	0.371+	0.365+	0.330+	0.385+	0.387+	<b>0.418</b>
Contraception	0.375	0.366+	0.392-	<b>0.398-</b>	0.393-	<b>0.398-</b>	<b>0.398-</b>	0.389-	0.375	0.375
Flags	0.332-	0.284+	0.316	0.3	<b>0.378-</b>	0.317	0.318-	0.348-	0.285+	0.306
Heart-c	0.702+	<b>0.812-</b>	0.634+	0.686+	0.716+	0.686+	0.691+	0.674+	0.754+	0.783
Heart-h	0.797	0.626+	0.658+	0.560+	0.644+	0.661+	0.602+	0.652+	0.763	<b>0.801</b>
Horse-colic	0.568	<b>0.587-</b>	0.586-	0.579	0.552	0.581	0.568	0.59	0.581-	0.562
Post-Operative	0.477	0.497-	0.463	0.436	0.457	0.47	0.463	0.510-	<b>0.525-</b>	0.451
Teaching-Assist-Eval	0.412	0.397+	0.412	0.414	0.425	0.414	0.414	0.400+	0.391+	<b>0.428</b>
Thoracic Surgery	0.635-	0.621-	<b>0.825-</b>	0.623-	0.640-	0.624-	0.634	0.558+	0.622-	0.571
Bridges-version-2	0.492	0.446+	<b>0.559-</b>	0.430+	0.430+	0.462	0.442+	0.482	0.488	0.498
Credit-approval	0.669+	0.524+	0.543+	0.537+	0.563+	0.572+	0.570+	<b>0.779</b>	0.620+	0.773
Credit-German	0.56	0.568	<b>0.663-</b>	0.556	0.56	0.557	0.579	0.540+	0.55	0.561
Cylinder-bands	<b>0.591-</b>	0.534+	0.582-	0.527+	0.536+	0.529+	0.509+	0.555	0.568	0.556
Dermatology	0.481+	0.485+	0.452+	0.431+	0.535	0.502+	0.504+	0.426+	0.383+	<b>0.538</b>
Heart-statlog	0.695+	<b>0.815-</b>	0.569+	0.643+	0.664+	0.630+	0.632+	0.704	0.743	0.734
Hepatitis	0.682	0.665	0.629+	0.672	0.624+	0.663	0.668	0.645	0.671	<b>0.696</b>
Labor	<b>0.7</b>	0.601	<b>0.7</b>	0.604	0.602	0.58	0.607	0.605	0.666	0.649
Liver-disorders	0.534+	0.555-	<b>0.566-</b>	0.544	0.549	0.544	0.544	0.558-	0.549	0.545
Tae	0.412	0.397+	0.412	0.414	0.411	0.414	0.414	0.409	0.391+	<b>0.417</b>
<b>Average</b>	<b>0.555</b>	<b>0.536</b>	<b>0.550</b>	<b>0.517</b>	<b>0.534</b>	<b>0.529</b>	<b>0.525</b>	<b>0.546</b>	<b>0.557</b>	<b>0.574</b>

Table 7: NMI results of  $k$ -prototypes on 15 synthetic mixed datasets.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
$S_1$	0.035+	0.003+	0.029+	0.008+	0.018+	0.008+	0.009+	0.018+	0.030+	<b>0.209</b>
$S_2$	0.421+	0.002+	0.446+	0.020+	0.007+	0.005+	0.004+	0.016+	0.043+	<b>0.557</b>
$S_3$	<b>0.459-</b>	0.004+	0.352	0.075+	0.014+	0.085+	0.013+	0.030+	0.003+	0.376
$S_4$	0.296	0.006+	0.258+	0.038+	0.027+	0.040+	0.048+	0.079+	0.007+	<b>0.332</b>
$S_5$	0.121+	0.010+	0.113+	0.034+	0.040+	0.035+	0.021+	0.031+	0.011+	<b>0.147</b>
$S_6$	0.256+	0.001+	0.204+	0.022+	0.014+	0.018+	0.019+	0.020+	0.000+	<b>0.464</b>
$S_7$	0.526	0.005+	0.386+	0.042+	0.047+	0.049+	0.029+	0.056+	0.006+	<b>0.587</b>
$S_8$	0.342+	0.013+	0.220+	0.051+	0.044+	0.042+	0.052+	0.038+	0.032+	<b>0.405</b>
$S_9$	0.351+	0.017+	0.284+	0.048+	0.053+	0.050+	0.056+	0.036+	0.017+	<b>0.447</b>
$S_{10}$	0.622+	0.005+	0.557+	0.058+	0.035+	0.042+	0.042+	0.096+	0.006+	<b>0.723</b>
$S_{11}$	0.204+	0.006+	0.178+	0.043+	0.028+	0.029+	0.042+	0.042+	0.035+	<b>0.329</b>
$S_{12}$	0.542	0.001+	0.467+	0.009+	0.014+	0.034+	0.004+	0.062+	0.035+	<b>0.579</b>
$S_{13}$	0.143+	0.034+	0.108+	0.057+	0.060+	0.055+	0.059+	0.077+	0.054+	<b>0.208</b>
$S_{14}$	0.180+	0.012+	0.113+	0.041+	0.038+	0.036+	0.041+	0.048+	0.087+	<b>0.289</b>
$S_{15}$	0.153+	0.001+	0.118+	0.008+	0.008+	0.006+	0.006+	0.039+	0.001+	<b>0.407</b>
<b>Average</b>	<b>0.310</b>	<b>0.008</b>	<b>0.255</b>	<b>0.037</b>	<b>0.030</b>	<b>0.036</b>	<b>0.030</b>	<b>0.046</b>	<b>0.025</b>	<b>0.404</b>

Table 8: Clustering Accuracy (ACC) results of  $k$ -prototypes on 15 synthetic mixed datasets.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
$S_1$	0.598+	0.528+	0.582+	0.546+	0.563+	0.543+	0.549+	0.549+	0.591+	<b>0.743</b>
$S_2$	0.843+	0.522+	0.863+	0.554+	0.539+	0.533+	0.529+	0.549+	0.587+	<b>0.9</b>
$S_3$	<b>0.779-</b>	0.361+	0.718	0.459+	0.387+	0.471+	0.375+	0.403+	0.360+	0.705
$S_4$	0.602	0.291+	0.563+	0.331+	0.323+	0.336+	0.333+	0.354+	0.287+	<b>0.606</b>
$S_5$	0.376	0.245+	0.369	0.278+	0.286+	0.281+	0.260+	0.268+	0.239+	<b>0.39</b>
$S_6$	0.754+	0.523+	0.703+	0.571+	0.559+	0.562+	0.541+	0.556+	0.510+	<b>0.853</b>
$S_7$	0.749	0.284+	0.628+	0.345+	0.345+	0.343+	0.316+	0.345+	0.285+	<b>0.784</b>
$S_8$	0.532+	0.212+	0.414+	0.258+	0.254+	0.253+	0.258+	0.243+	0.229+	<b>0.586</b>
$S_9$	0.503+	0.194+	0.461+	0.230+	0.236+	0.232+	0.235+	0.218+	0.193+	<b>0.595</b>
$S_{10}$	0.794	0.281+	0.768+	0.359+	0.340+	0.346+	0.336+	0.385+	0.284+	<b>0.857</b>
$S_{11}$	0.488+	0.288+	0.459+	0.336+	0.328+	0.325+	0.331+	0.341+	0.319+	<b>0.587</b>
$S_{12}$	0.891	0.521+	0.859	0.543+	0.548+	0.568+	0.530+	0.618+	0.571+	<b>0.904</b>
$S_{13}$	0.261+	0.158+	0.224+	0.181+	0.184+	0.181+	0.180+	0.191+	0.180+	<b>0.327</b>
$S_{14}$	0.376+	0.209+	0.326+	0.249+	0.245+	0.243+	0.245+	0.253+	0.296+	<b>0.461</b>
$S_{15}$	0.698+	0.510+	0.671+	0.536+	0.542+	0.529+	0.532+	0.586+	0.514+	<b>0.832</b>
<b>Average</b>	<b>0.616</b>	<b>0.342</b>	<b>0.574</b>	<b>0.385</b>	<b>0.378</b>	<b>0.383</b>	<b>0.370</b>	<b>0.391</b>	<b>0.363</b>	<b>0.675</b>

Table 9: NMI results of EM on 20 mixed datasets from the UCI repository.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
Acute-Inflamations	0.025+	0.018+	0.141+	0.036+	0.036+	0.036+	0.036+	0.188+	0.414	<b>0.447</b>
Automobile	0.173+	0.158+	0.164+	0.177+	0.198+	0.186+	0.156+	0.244+	0.178+	<b>0.299</b>
Contraception	<b>0.03</b>	0.019+	0.015+	<b>0.03</b>	<b>0.03</b>	<b>0.03</b>	<b>0.03</b>	0.022	<b>0.03</b>	<b>0.03</b>
Flags	0.161	0.142	0.147	0.128	<b>0.203-</b>	0.133	0.120+	0.136	0.127+	0.146
Heart-c	0.122+	0.226+	0.130+	0.044+	0.097+	0.043+	0.048+	0.109+	0.149+	<b>0.3</b>
Heart-h	0.273	0.006+	0.099+	0.036+	0.124+	0.091+	0.013+	0.100+	0.240+	<b>0.313</b>
Horse-colic	0.053	0.001+	0.032+	0.004+	0.037	0.004+	0.003+	0.013+	<b>0.091-</b>	0.047
Post-Operative	0.028	0.02	0.031	0.013+	0.02	0.024	0.026	<b>0.037</b>	0.025	0.021
Teaching-Assist-Eval	<b>0.051-</b>	0.026-	0.048-	0.047-	0.015	0.047-	0.047-	0.028-	0.023-	0.015
Thoracic Surgery	0.005+	0.005+	0.003+	0.004+	0.002+	0.003+	0.003+	<b>0.012</b>	0.003	0.008
Bridges-version-2	0.353	0.294+	<b>0.36</b>	0.219+	0.119+	0.143+	0.118+	0.265+	0.326+	0.343
Credit-approval	0.025+	0.021+	0.025+	0.065+	0.074+	0.070+	0.021+	0.029+	0.033+	<b>0.265</b>
Credit-german	<b>0.019-</b>	0	0.000+	0.005	0.002	0.005	0.004	0.004	0.005	0.003
Cylinder-bands	0.005	0.009	0.005	0.003+	0.014	0.004+	0.004	<b>0.028-</b>	0.01	0.004
Dermatology	0.305+	0.418+	0.324+	0.469+	<b>0.652</b>	0.592+	0.597+	0.351+	0.294+	0.642
Heart-statlog	0.109+	0.216+	0.047+	0.053+	0.098+	0.045+	0.078+	0.123+	0.119+	<b>0.277</b>
Hepatitis	<b>0.144-</b>	0.085+	0.076+	0.049+	0.070+	0.048+	0.029+	0.029+	0.108	0.108
Labor	0.201	0.136+	<b>0.23</b>	0.089+	0.085+	0.043+	0.058+	0.037+	0.102+	0.221
Liver-disorders	0.008+	0.004+	0.003+	0.003+	0.008+	0.003+	0.003+	0.004+	0.01	<b>0.016</b>
Tae	<b>0.051</b>	0.026+	0.047	0.047	0.027+	0.047	0.047	0.039	0.022+	<b>0.051</b>
<b>Average</b>	<b>0.107</b>	<b>0.092</b>	<b>0.096</b>	<b>0.076</b>	<b>0.096</b>	<b>0.080</b>	<b>0.072</b>	<b>0.090</b>	<b>0.115</b>	<b>0.178</b>

Table 10: Clustering Accuracy (ACC) results of EM on 20 mixed datasets from the UCI repository.

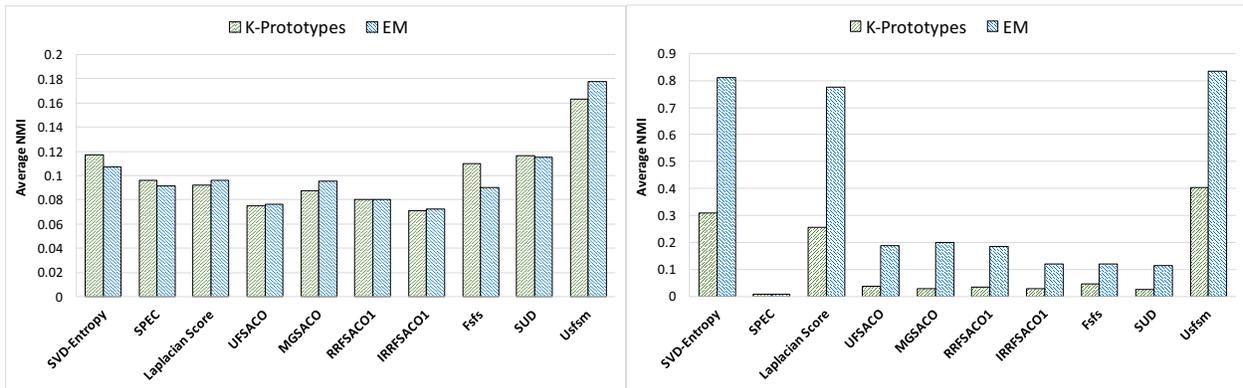
Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
Acute-Inflamations	0.583+	0.575+	0.655+	0.608+	0.608+	0.608+	0.608+	0.711	<b>0.836</b>	0.825
Automobile	0.387+	0.390+	0.376+	0.379+	0.388+	0.380+	0.348+	0.417+	0.384+	<b>0.452</b>
Contraception	0.377	0.373	0.392-	0.395-	<b>0.401-</b>	0.395-	0.395-	0.396-	0.377	0.377
Flags	0.305	0.295	0.292	0.318-	<b>0.375-</b>	0.304	0.325-	0.363-	0.302	0.288
Heart-c	0.686+	0.769+	0.698+	0.596+	0.659+	0.588+	0.592+	0.663+	0.702+	<b>0.812</b>
Heart-h	0.803	0.639+	0.709+	0.564+	0.677+	0.673+	0.643+	0.666+	0.799+	<b>0.833</b>
Horse-colic	0.585	0.545+	0.522+	0.583	0.588	0.584	0.591-	0.594-	<b>0.636-</b>	0.577
Post-Operative	0.446	0.396+	0.463	0.419	0.44	0.451	0.45	0.506	<b>0.516-</b>	0.43
Teaching-Assist-Eval	0.412-	0.395-	0.412-	<b>0.414-</b>	0.344	<b>0.414-</b>	<b>0.414-</b>	0.377-	0.381-	0.344
Thoracic Surgery	0.634-	0.813-	<b>0.824-</b>	0.738-	0.742-	0.691	0.823-	0.563+	0.808-	0.604
Bridges-version-2	<b>0.515-</b>	0.444+	0.562-	0.436+	0.430+	0.473	0.447	0.478	0.477	0.482
Credit-approval	0.548+	0.580+	0.549+	0.632+	0.608+	0.639+	0.579+	0.590+	0.553+	<b>0.769</b>
Credit-german	0.632-	0.586-	<b>0.637-</b>	0.595-	0.574	0.585-	0.598-	0.538	0.554	0.532
Cylinder-bands	0.553-	0.557-	0.556-	0.544	0.57	0.541	0.530+	<b>0.601-</b>	0.571-	0.54
Dermatology	0.486+	0.554+	0.490+	0.532+	<b>0.717-</b>	0.633	0.633	0.441+	0.410+	0.662
Heart-statlog	0.672+	0.765+	0.601+	0.609+	0.660+	0.597+	0.632+	0.679+	0.671+	<b>0.802</b>
Hepatitis	0.675+	0.664+	0.553+	0.689	0.669	0.691	0.668	0.663	<b>0.708</b>	0.69
Labor	0.725	0.643+	0.758	0.612+	0.629+	0.571+	0.590+	0.625+	0.706+	<b>0.77</b>
Liver-disorders	0.529	0.564-	0.566-	<b>0.568-</b>	0.546-	<b>0.568-</b>	<b>0.568-</b>	0.566-	0.53	0.526
Tae	0.412	0.395+	0.412	0.414	0.404	0.414	0.414	0.405	0.395+	<b>0.417</b>
<b>Average</b>	<b>0.548</b>	<b>0.547</b>	<b>0.551</b>	<b>0.532</b>	<b>0.551</b>	<b>0.540</b>	<b>0.542</b>	<b>0.542</b>	<b>0.566</b>	<b>0.587</b>

Table 11: NMI results of EM on 15 synthetic mixed datasets.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
$S_1$	0.122+	0.001+	0.082+	0.043+	0.086+	0.043+	0.036+	0.038+	0.158+	<b>0.438</b>
$S_2$	0.751	0.001+	0.742	0.079+	0.071+	0.026+	0.023+	0.042+	0.226+	<b>0.761</b>
$S_3$	<b>0.877-</b>	0.003+	0.815	0.167+	0.097+	0.188+	0.031+	0.046+	0.003+	0.797
$S_4$	<b>0.813-</b>	0.007+	0.767	0.157+	0.193+	0.124+	0.060+	0.161+	0.008+	0.723
$S_5$	<b>0.73</b>	0.011+	0.728	0.118+	0.127+	0.118+	0.034+	0.089+	0.009+	0.656
$S_6$	0.801+	0.001+	0.833+	0.232+	0.298+	0.264+	0.057+	0.077+	0.001+	<b>0.882</b>
$S_7$	<b>0.983</b>	0.005+	0.945+	0.176+	0.248+	0.190+	0.032+	0.128+	0.005+	0.98
$S_8$	0.924	0.015+	0.899	0.179+	0.108+	0.140+	0.067+	0.079+	0.095+	<b>0.933</b>
$S_9$	0.9	0.019+	0.893	0.097+	0.151+	0.108+	0.111+	0.061+	0.017+	<b>0.921</b>
$S_{10}$	0.999	0.004+	<b>1</b>	0.473+	0.441+	0.454+	0.337+	0.202+	0.006+	<b>1</b>
$S_{11}$	<b>0.845</b>	0.006+	0.766+	0.288+	0.308+	0.315+	0.243+	0.112+	0.120+	0.843
$S_{12}$	0.779	0.001+	0.791	0.081+	0.118+	0.115+	0.023+	0.229+	0.352+	<b>0.81</b>
$S_{13}$	0.79	0.033+	0.555+	0.100+	0.088+	0.092+	0.100+	0.102+	0.192+	<b>0.817</b>
$S_{14}$	0.911+	0.010+	0.891+	0.235+	0.180+	0.190+	0.320+	0.096+	0.534+	<b>0.957</b>
$S_{15}$	0.937+	0.001+	0.937+	0.386+	0.473+	0.404+	0.348+	0.326+	0.002+	<b>1</b>
<b>Average</b>	<b>0.811</b>	<b>0.008</b>	<b>0.776</b>	<b>0.187</b>	<b>0.199</b>	<b>0.185</b>	<b>0.121</b>	<b>0.119</b>	<b>0.115</b>	<b>0.834</b>

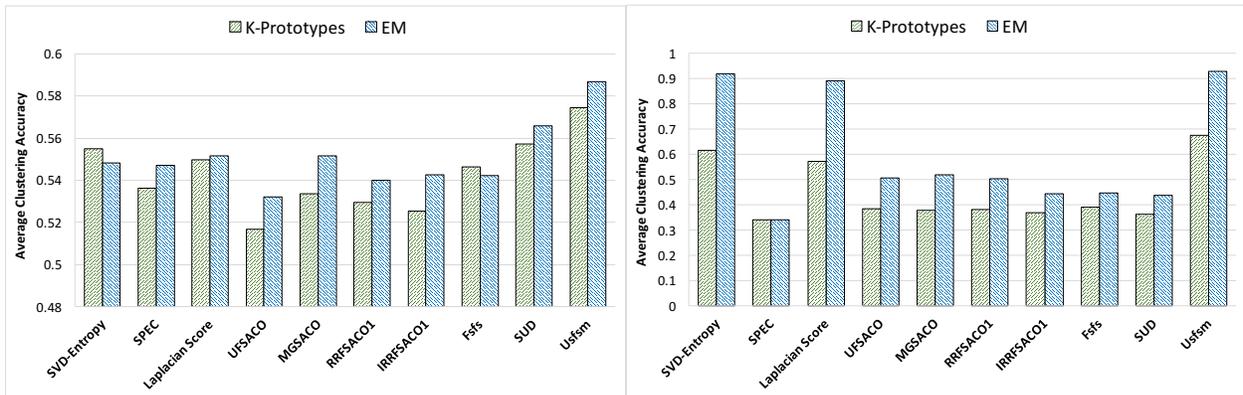
Table 12: Clustering Accuracy (ACC) results of EM on 15 synthetic mixed datasets.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
$S_1$	0.672+	0.511+	0.620+	0.585+	0.640+	0.587+	0.572+	0.565+	0.714+	<b>0.866</b>
$S_2$	0.959	0.518+	0.957	0.610+	0.618+	0.568+	0.570+	0.588+	0.714+	<b>0.961</b>
$S_3$	<b>0.971-</b>	0.358+	0.953	0.518+	0.475+	0.554+	0.393+	0.416+	0.361+	0.948
$S_4$	<b>0.943-</b>	0.286+	0.895	0.427+	0.448+	0.403+	0.340+	0.432+	0.297+	0.906
$S_5$	<b>0.834</b>	0.252+	0.838	0.329+	0.343+	0.340+	0.272+	0.303+	0.252+	0.829
$S_6$	0.970+	0.518+	0.976+	0.733+	0.766+	0.739+	0.586+	0.622+	0.520+	<b>0.984</b>
$S_7$	<b>0.995</b>	0.284+	0.986+	0.454+	0.526+	0.469+	0.317+	0.425+	0.286+	<b>0.995</b>
$S_8$	0.93	0.214+	0.925	0.377+	0.321+	0.350+	0.277+	0.287+	0.276+	<b>0.966</b>
$S_9$	0.929	0.194+	0.918	0.262+	0.319+	0.276+	0.267+	0.230+	0.191+	<b>0.943</b>
$S_{10}$	1	0.280+	1	0.673+	0.677+	0.670+	0.542+	0.462+	0.293+	<b>1</b>
$S_{11}$	<b>0.900-</b>	0.284+	0.795	0.556+	0.555+	0.571+	0.499+	0.409+	0.389+	0.806
$S_{12}$	0.964	0.518+	0.966	0.607+	0.651+	0.636+	0.555+	0.724+	0.809+	<b>0.971</b>
$S_{13}$	0.801	0.158+	0.611+	0.211+	0.205+	0.209+	0.211+	0.214+	0.294+	<b>0.824</b>
$S_{14}$	0.912	0.207+	0.919	0.421+	0.381+	0.381+	0.456+	0.287+	0.621+	<b>0.947</b>
$S_{15}$	0.993+	0.527+	0.993+	0.830+	0.869+	0.825+	0.794+	0.764+	0.579+	<b>1</b>
<b>Average</b>	<b>0.918</b>	<b>0.341</b>	<b>0.890</b>	<b>0.506</b>	<b>0.520</b>	<b>0.505</b>	<b>0.443</b>	<b>0.449</b>	<b>0.440</b>	<b>0.930</b>



(a) Average NMI on real datasets

(b) Average ACC on real datasets



(c) Average NMI on synthetic datasets

(d) Average ACC on synthetic datasets

Figure 9: Average NMI and ACC of  $k$ -prototypes and EM over the real and synthetic datasets.

average to all the other methods including SUD, an unsupervised feature selection method for mixed data. Moreover, our method has a statistical significant better behavior than most of the other methods. Likewise, for synthetic datasets, in Tables 7-8 we can see that the proposed method got the best performance on average among all competing methods, and it had a statistical significant better behavior than almost all feature selection methods in all datasets, except for the  $S_3$  dataset, where it was the second best. In these datasets, the second and third best methods were SVD-Entropy and Laplacian score respectively. On the other hand, for EM, Tables 9-12 show that the results obtained using this clustering algorithm are very similar to  $k$ -prototypes, but in general, with a higher clustering quality, especially regarding ACC. A comparative chart of the results obtained by  $k$ -prototypes and EM is shown in Figure 9. In Figures 9a-9d we can see the average results regarding NMI and ACC over the real and synthetic datasets. In these figures, it is possible to observe that our method obtained the best results for both clustering algorithms.

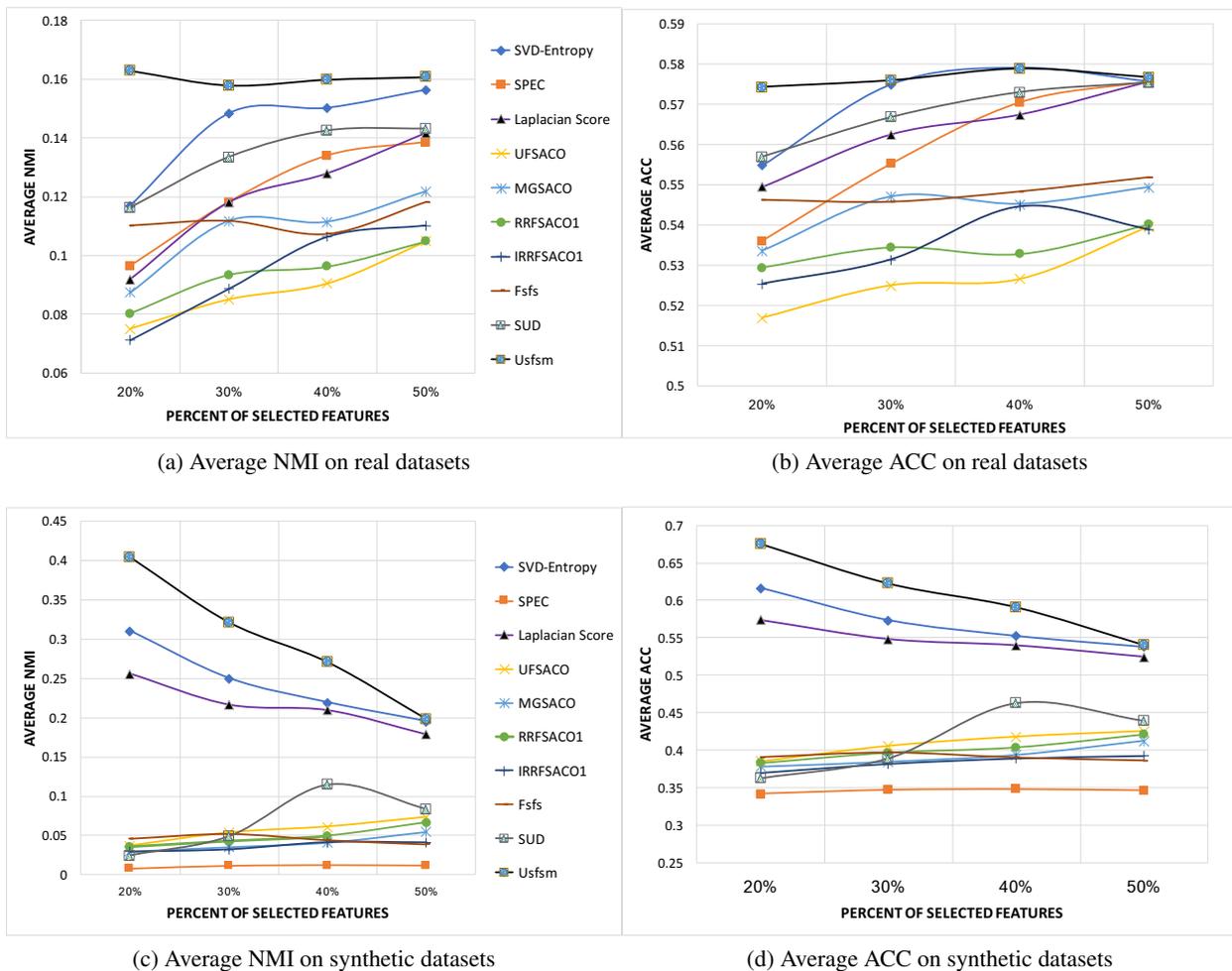


Figure 10: Average NMI and ACC of  $k$ -prototypes selecting the 20-50% of features over the real and synthetic datasets.

Furthermore, in order to show the performance of the unsupervised feature selection methods using a greater number of features to be selected from each dataset. We tested 30%, 40% and 50% of the ranked

features for methods based on ranking, and the same percentages of the whole set of features for feature subset selection methods. In Figures 10a and 10b, the average results of  $k$ -prototypes regarding NMI and ACC over the real datasets are shown respectively. In these figures, we can see that our method outperforms the other methods, especially selecting 20% of the features of each dataset. This indicates that USFSM is able to place the most relevant features at the beginning of the ranking. It is also appreciated in these figures that the SVD-Entropy and SUD methods were the second and the third best methods respectively. In the same way, the results for the synthetic datasets (see Figures 10c and 10d) show a similar result, except that, in this case, the third best method is the Laplacian score. Note that USFSM, SVD-Entropy, and Laplacian score methods are significantly better than the remaining methods on these synthetic datasets. Similar results were obtained with EM (see Figure 11).

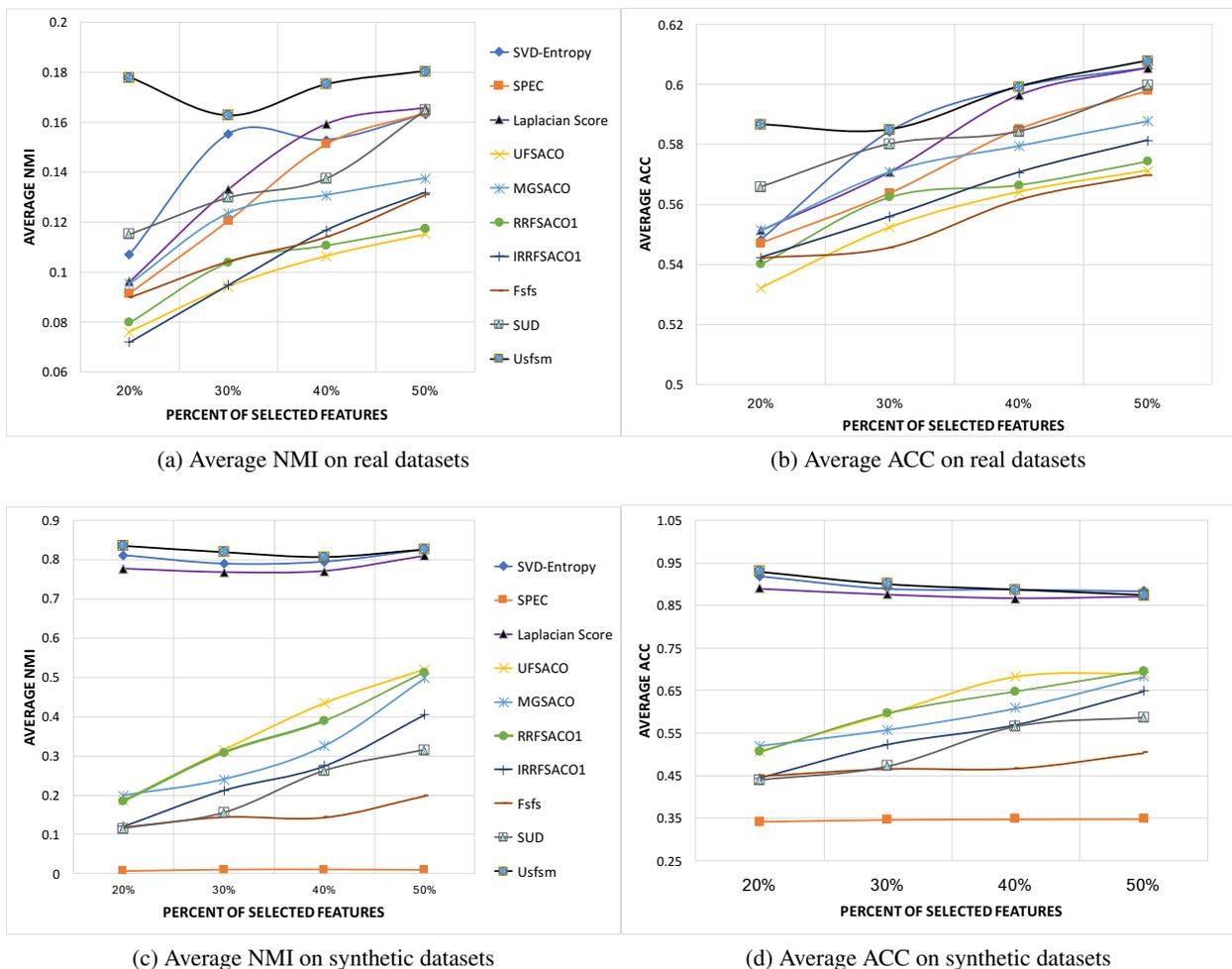


Figure 11: Average NMI and ACC of EM selecting the 20-50% of features over the real and synthetic datasets.

From the results of Tables 5-12 and Figures 9-11, it can be concluded that USFSM outperforms to SUD in terms ACC and NMI using  $k$ -prototypes and EM, indicating that feature discretization in general provides worse results in these datasets. Moreover, our method proved to be better than the unsupervised filter meth-

ods SVD-Entropy, SPEC, Laplacian score, UFSACO, MGSACO, RRFSAO1, IRRFSACO1, and Fsf; all of them designed for numerical data, which gives us evidence that the coding of features (necessary for the application of these methods) produces in most cases worse results. Furthermore, the results obtained in this experiment show the ability of our method to set the most relevant features at the beginning of the ranking.

## 6.2.2 Evaluation in terms of supervised classification

In this experiment, we evaluate the proposed method in terms of supervised classification results using the previously mentioned classifiers and the evaluation strategy described in Section 2.3. Feature evaluation was made using the same percentage of features to select as in the previous experiment, that is, 20%.

Table 13: Classification accuracy of SVM on 20 mixed datasets from the UCI repository.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsf	SUD	Ufsm
Acute-Inflamations	0.583+	0.575+	0.658+	0.625+	0.625+	0.625+	0.625+	0.742	0.767	<b>0.825</b>
Automobile	0.438	0.424+	0.429	0.405+	0.488	0.46	0.469	<b>0.624-</b>	0.439	0.502
Contraception	0.449	0.427+	0.420+	<b>0.481</b>	0.465	<b>0.481</b>	<b>0.481</b>	0.436	0.449	0.449
Flags	<b>0.444</b>	0.331	0.387	0.407	0.634-	0.448	0.382	0.397	0.309	0.362
Heart-c	0.693	<b>0.788-</b>	0.72	0.726	0.73	0.743	0.719	0.7	0.746	0.739
Heart-h	0.813	0.646+	0.684+	0.663+	0.731	0.724	0.663+	0.785	0.813	<b>0.817</b>
Horse-colic	0.639	0.611	0.707	0.688	<b>0.718</b>	0.677	0.611	0.704	0.715	0.655
Post-Operative	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	<b>0.7</b>	<b>0.711</b>	<b>0.711</b>
Teaching-Assist-Eval	0.398+	0.318+	0.397+	0.384+	0.497	0.384+	0.384+	0.390+	0.383+	<b>0.51</b>
Thoracic Surgery	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	0.849	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>
Bridges-version-2	0.602	0.45	<b>0.619-</b>	0.475	0.455+	0.474	0.475	0.497	0.507	0.535
Credit-approval	0.730+	0.561+	0.567+	0.658+	0.659+	0.683+	0.612+	0.835	0.655+	<b>0.858</b>
Credit-german	0.702	0.7	0.7	0.701	0.7	0.701	0.701	0.701	<b>0.717-</b>	0.703
Cylinder-bands	0.661	0.606+	0.652	0.630+	<b>0.744-</b>	0.635+	0.576+	0.633+	0.630+	0.687
Dermatology	0.642+	0.689+	0.653+	0.669+	0.803	0.814-	<b>0.874-</b>	0.650+	0.497+	0.754
Heart-statlog	0.693	0.793	0.607+	0.707	0.678	0.715	0.704	<b>0.759</b>	0.741	0.73
Hepatitis	0.794	0.794	0.844-	0.813	0.787	0.813	0.8	0.787	<b>0.845</b>	0.787
Labor	0.847	0.753	0.823	0.773	0.79	0.77	0.81	0.777	0.77	<b>0.863</b>
Liver-disorders	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>
Tae	<b>0.398</b>	0.318	0.397	0.384	0.358	0.384	0.384	0.364	0.332	0.351
<b>Average</b>	<b>0.633</b>	<b>0.596</b>	<b>0.620</b>	<b>0.617</b>	<b>0.650</b>	<b>0.634</b>	<b>0.621</b>	<b>0.646</b>	<b>0.623</b>	<b>0.663</b>

Table 14: Classification accuracy of KNN ( $K = 3$ ) on 20 mixed datasets from the UCI repository.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsf	SUD	Ufsm
Acute-Inflamations	0.583+	0.575+	0.658+	0.550+	0.550+	0.550+	0.550+	0.7	0.767	<b>0.825</b>
Automobile	0.65	0.605	0.605	0.693-	0.678-	0.697-	<b>0.732-</b>	0.633	0.644-	0.553
Contraception	0.461	0.419+	0.425+	<b>0.486</b>	0.453	<b>0.486</b>	<b>0.486</b>	0.431	0.461	0.461
Flags	0.453	0.367	0.397	0.401	<b>0.592-</b>	0.427	0.36	0.386	0.283	0.376
Heart-c	0.651+	0.792	0.720+	0.627+	0.614+	0.643+	0.624+	0.664+	0.7	<b>0.796</b>
Heart-h	0.741+	0.650+	0.687+	0.606+	0.687+	0.697+	0.578+	0.789	<b>0.83</b>	0.823
Horse-colic	0.668	0.657	0.68	0.668	0.652	0.647	0.642	0.709	<b>0.739-</b>	0.661
Post-Operative	0.678	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	0.667	0.667	<b>0.711</b>
Teaching-Assist-Eval	0.398	0.431	0.397	0.384	<b>0.457</b>	0.384	0.384	0.397	0.443	<b>0.457</b>
Thoracic Surgery	0.845	0.817+	0.791+	0.768+	0.760+	0.764+	0.766+	<b>0.857</b>	0.847	0.849
Bridges-version-2	0.611	0.459	<b>0.62</b>	0.457	0.495	0.455	0.455	0.479	0.487	0.535
Credit-approval	0.735+	0.649+	0.583+	0.655+	0.639+	0.632+	0.617+	0.799+	0.699+	<b>0.867</b>
Credit-german	0.649+	0.701	0.68	0.651+	0.676	0.651+	0.664	<b>0.702</b>	0.69	0.692
Cylinder-bands	0.667	0.681	<b>0.761-</b>	0.661	0.737-	0.665	0.604+	0.661	0.720-	0.661
Dermatology	0.642+	0.653+	0.647+	0.625	0.773	0.787-	<b>0.806-</b>	0.628+	0.495+	0.748
Heart-statlog	0.633+	<b>0.811</b>	0.615+	0.585+	0.596+	0.596+	0.581+	0.726	0.663+	0.752
Hepatitis	0.748	0.710+	0.831	0.735	0.718+	0.742	0.705+	0.697+	0.813	<b>0.826</b>
Labor	<b>0.897</b>	0.823	0.863	0.773	0.83	0.813	0.757	0.817	0.79	0.847
Liver-disorders	0.501	0.533	<b>0.582</b>	0.478+	0.574	0.478+	0.478+	0.484	0.574	0.568
Tae	0.398	0.431	0.397	0.384	0.437	0.384	0.384	0.391	<b>0.437-</b>	0.351
<b>Average</b>	<b>0.630</b>	<b>0.624</b>	<b>0.633</b>	<b>0.595</b>	<b>0.631</b>	<b>0.610</b>	<b>0.594</b>	<b>0.631</b>	<b>0.637</b>	<b>0.668</b>

Table 15: Classification accuracy of NB on 20 mixed datasets from the UCI repository.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
Acute-Inflamations	0.583+	0.575+	0.658+	0.483+	0.483+	0.483+	0.483+	0.683	0.767	<b>0.825</b>
Automobile	0.532	0.366+	0.424+	0.484	0.58	0.508	<b>0.576</b>	0.57	0.474	0.512
Contraception	0.445	0.407+	0.428	<b>0.482</b>	0.471	<b>0.482</b>	<b>0.482</b>	0.437	0.445	0.445
Flags	0.469	0.336	0.403	0.431	<b>0.609-</b>	0.443	0.356	0.444	0.325	0.398
Heart-c	0.723+	0.810	0.720+	0.729	0.746	0.746	0.73	0.697+	0.729+	<b>0.812</b>
Heart-h	<b>0.82</b>	0.650+	0.687+	0.660+	0.752	0.707+	0.657+	0.782	0.813	<b>0.82</b>
Horse-colic	0.726	0.638	0.704	0.702	0.717	0.669	0.641	0.718	<b>0.731</b>	0.696
Post-Operative	0.711	<b>0.722</b>	0.7	0.711	0.711	0.711	0.711	0.689	<b>0.722</b>	0.711
Teaching-Assist-Eval	0.398	0.344+	0.397	0.384	0.45	0.384	0.384	0.397	0.409	<b>0.457</b>
Thoracic Surgery	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>	<b>0.851</b>
Bridges-version-2	0.593	0.478	<b>0.601</b>	0.485	0.476	0.455+	0.445+	0.497	0.507	0.535
Credit-approval	0.717+	0.664+	0.583+	0.728+	0.713+	0.713+	0.675+	0.833	0.728+	<b>0.864</b>
Credit-german	0.673	0.7	0.699	0.7	0.701	0.708	0.709	0.712	<b>0.720-</b>	0.7
Cylinder-bands	<b>0.707</b>	0.7	0.63	0.694	0.761	0.698	0.659	0.606+	0.676	0.678
Dermatology	0.642+	0.721+	0.647+	0.682	0.819	0.814-	<b>0.858-</b>	0.653+	0.495+	0.77
Heart-statlog	0.726	<b>0.819</b>	0.648+	0.726+	0.681+	0.744	0.726	0.77	0.733	0.793
Hepatitis	0.82	0.813	0.837	0.826	0.794	0.832	0.781	0.787	<b>0.852</b>	0.792
Labor	0.813	0.807	0.81	0.7	0.81	0.72	0.74	0.797	0.77	<b>0.863</b>
Liver-disorders	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>	<b>0.58</b>
Tae	<b>0.398</b>	0.344	0.397	0.384	0.344	0.384	0.384	0.371	0.344	0.351
<b>Average</b>	<b>0.646</b>	<b>0.617</b>	<b>0.620</b>	<b>0.621</b>	<b>0.653</b>	<b>0.632</b>	<b>0.621</b>	<b>0.644</b>	<b>0.634</b>	<b>0.672</b>

Table 16: Classification accuracy of Random-Forest on 20 mixed datasets from the UCI repository.

Dataset	SVD-Entropy	SPEC	Laplacian Score	UFSACO	MGSACO	RRFSACO1	IRRFSAO1	Fsfs	SUD	Ufsm
Acute-Inflamations	0.583+	0.575+	0.658+	0.533+	0.533+	0.533+	0.533+	0.692	0.767	<b>0.825</b>
Automobile	<b>0.722-</b>	0.571	0.634	0.673-	0.696-	0.712-	0.752-	0.633	0.658-	0.552
Contraception	0.461	0.419+	0.424+	<b>0.492</b>	0.451	<b>0.492</b>	<b>0.492</b>	0.433	0.461	0.461
Flags	<b>0.479</b>	0.377	0.412	0.381	0.608-	0.376	0.346	0.386	0.35	0.36
Heart-c	0.637+	<b>0.809</b>	0.72	0.647+	0.631+	0.647+	0.637+	0.67	0.683+	0.772
Heart-h	0.762	0.650+	0.684+	0.612+	0.680+	0.694+	0.619+	0.789	0.83	<b>0.833</b>
Horse-colic	0.641	0.658	0.682	0.671	0.669	0.66	0.614+	0.726	<b>0.736</b>	0.682
Post-Operative	0.678	<b>0.711</b>	<b>0.711</b>	0.678	<b>0.711</b>	<b>0.711</b>	<b>0.711</b>	0.667	0.667	<b>0.711</b>
Teaching-Assist-Eval	0.398+	0.471	0.397	0.384+	0.47	0.384+	0.384+	0.424	<b>0.49</b>	<b>0.49</b>
Thoracic Surgery	0.845	0.832	0.804+	0.798+	0.798+	0.809+	0.815+	0.85	0.845	<b>0.853</b>
Bridges-version-2	0.612	0.432+	<b>0.629</b>	0.402+	0.418+	0.418+	0.418+	0.46	0.488	0.545
Credit-approval	0.735+	0.686+	0.583+	0.667+	0.668+	0.686+	0.652+	0.801	0.697+	<b>0.867</b>
Credit-german	0.662	0.701-	0.674	0.636+	0.647+	0.655	0.648	<b>0.707</b>	0.685	0.684
Cylinder-bands	0.68	0.72	<b>0.778-</b>	0.696	0.652	0.713	0.739	0.676	0.728	0.687
Dermatology	0.639+	0.639+	0.639+	0.611+	0.809-	<b>0.759</b>	0.828-	0.631+	0.492+	0.74
Heart-statlog	0.685+	<b>0.822-</b>	0.615+	0.656+	0.670+	0.656+	0.626+	0.744	0.7	0.752
Hepatitis	0.801	0.821	0.831	0.782	0.769	0.78	0.722	0.736	<b>0.833</b>	0.826
Labor	<b>0.88</b>	0.823	0.813	0.69	0.833	0.78	0.827	0.76	0.79	0.843
Liver-disorders	0.524+	0.533+	0.597	0.469+	0.577	0.469+	0.469+	0.502+	0.6	<b>0.603</b>
Tae	0.398	0.471	0.397	0.384	<b>0.477-</b>	0.384	0.384	0.418	0.470-	0.351
<b>Average</b>	<b>0.641</b>	<b>0.636</b>	<b>0.634</b>	<b>0.593</b>	<b>0.638</b>	<b>0.616</b>	<b>0.611</b>	<b>0.636</b>	<b>0.649</b>	<b>0.671</b>

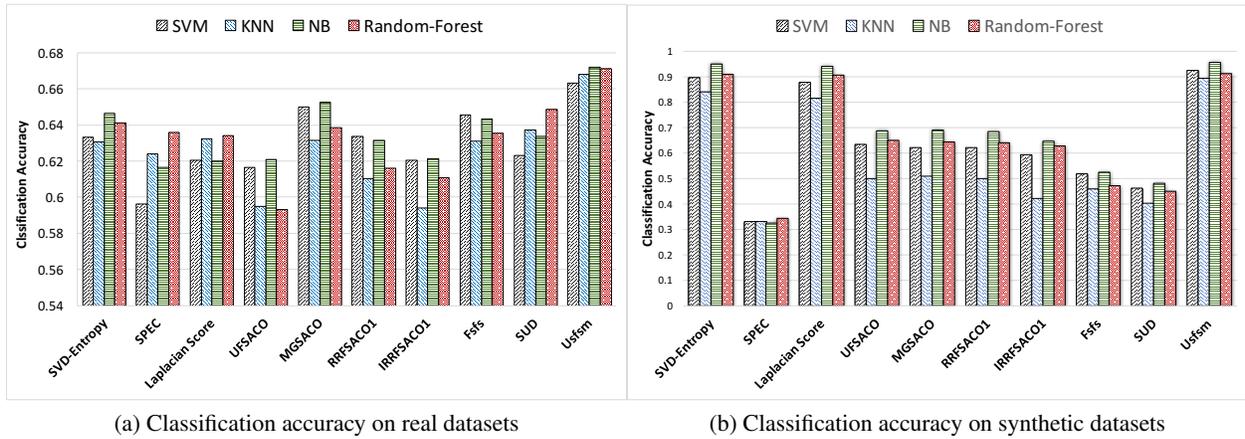


Figure 12: Average classification performance of SVM, KNN ( $K = 3$ ), NB, and Random Forest on real and synthetic datasets.

Tables 13-16 show the classification results over the real datasets of Table 3 using SVM, KNN ( $K = 3$ ), Naive Bayes (NB), and Random-Forest, respectively. As it can be seen in Table 13, our method overcomes all other methods on average, using SVM. Moreover, our method showed a significantly better performance than most other methods in many datasets. A similar result is shown in Table 14 for KNN ( $K = 3$ ), where we can see that USFSM outperforms all other methods on average, and it is significantly better than most other methods. Likewise, Table 15, shows the classification results using Naive Bayes. With this classifier, our method obtains the best results on average among the used classifiers, outperforming the other unsupervised feature selection methods, and achieving a significantly better behavior. Meanwhile, using Random-Forest (see Table 16) our method again got the best results on average, having a statistically better behavior. On the other hand, for the synthetic datasets, the results were very similar to those obtained in the real datasets. A comparative chart of the average classification accuracy achieved by the four classifiers over the real and synthetic datasets of Tables 3-4 is shown in Figure 12, where the superiority of the proposed method can be appreciated.

Finally, we also show the performance of the classifiers using different percentages of selected features (i.e., 20%, 30%, 40% and 50%, as in the first experiment). In Figure 13, the average classification results obtained by SVM, KNN ( $K = 3$ ), NB, and Random-Forest over the real datasets can be observed. The results showed in Figures 13a-13d, provide evidence that the proposed method is the best on average in the real datasets. In specific, using SVM and NB, we can see that our method get the best results in all percentages of selected features. For KNN ( $K = 3$ ) and Random-Forest, it can be observed that using 20% and 30% of the features it is possible to obtain the best results. With higher percentages, the Laplacian score, SVD-Entropy, and SUD methods yield slightly better results than ours. Likewise, for the synthetic data, in Figures 14a-14d, it is observed that the best results are achieved by our method, being SVD-Entropy and Laplacian Score, the second and third best methods respectively. Note that in these synthetic datasets, USFSM, SVD-Entropy, and the Laplacian score are significantly better than the rest.

From the results shown in Tables 13-16 as well as of Figures 12-14, it can be concluded that our method achieved the best classification results over real and synthetic datasets compared to other unsupervised feature selection methods of the state-of-the-art. Especially with SVM and NB, where it always showed the best performance. It is also possible to appreciate again through the results that our method places the

most relevant features at the beginning of the ranking, and that both the feature discretization (used in SUD), as well as feature codification employed for the unsupervised feature selection methods for numerical data, in general, give worse results.

### 6.3 Discussion

In this section, the preliminary results obtained so far in this Ph.D. research were presented. We have introduced a new unsupervised feature selection method for mixed data called USFSM. Experiments for evaluating and comparing USFSM against other state-of-the-art methods were carried out. The performance of our method in terms of clustering results using two popular clustering algorithms, as well as the performance of the proposed method in terms of the classification quality using several classifiers were conducted. Besides the better results of the proposed method, from our experiments, we can conclude the following:

1. The experimental results provide evidence that it is possible to identify the relevant features in mixed

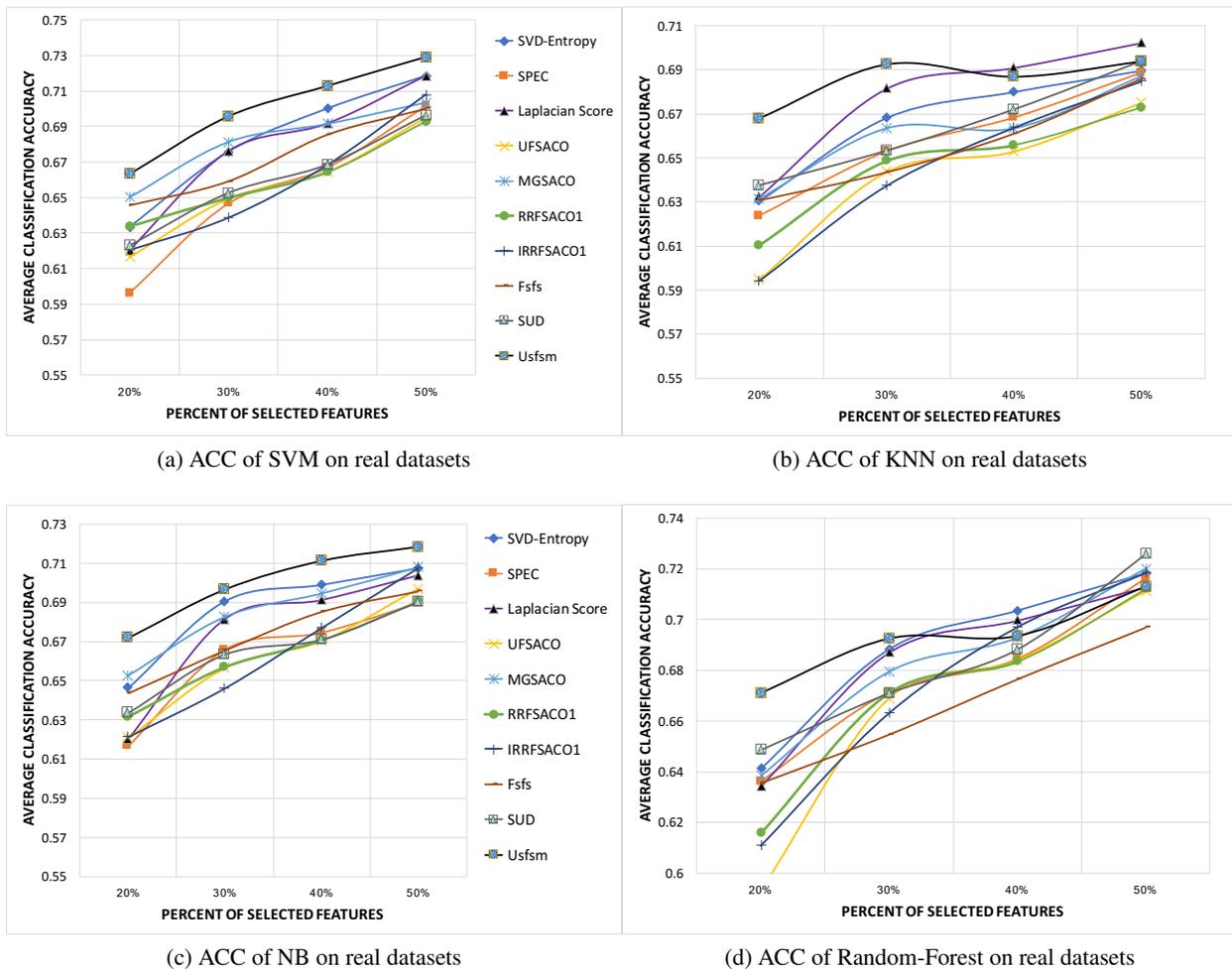


Figure 13: Average classification accuracy of SVM, KNN, NB and Random-Forest selecting the 20-50% of features over the real datasets.

datasets.

2. Feature discretization or feature encoding as a solution for unsupervised feature selection under mixed data, in general, produce worse results; this suggests that avoiding such procedures may help to achieve better results.
3. Spectral feature selection theory can be used for identifying relevant features in mixed data.

## 7 Conclusions

This Ph.D. research proposal is focused on the problem of unsupervised feature selection for mixed data. We present a revision of the related work to show the most relevant approaches that have been proposed to solve this problem. Based on this review we highlight the need for further research in this area. Then,

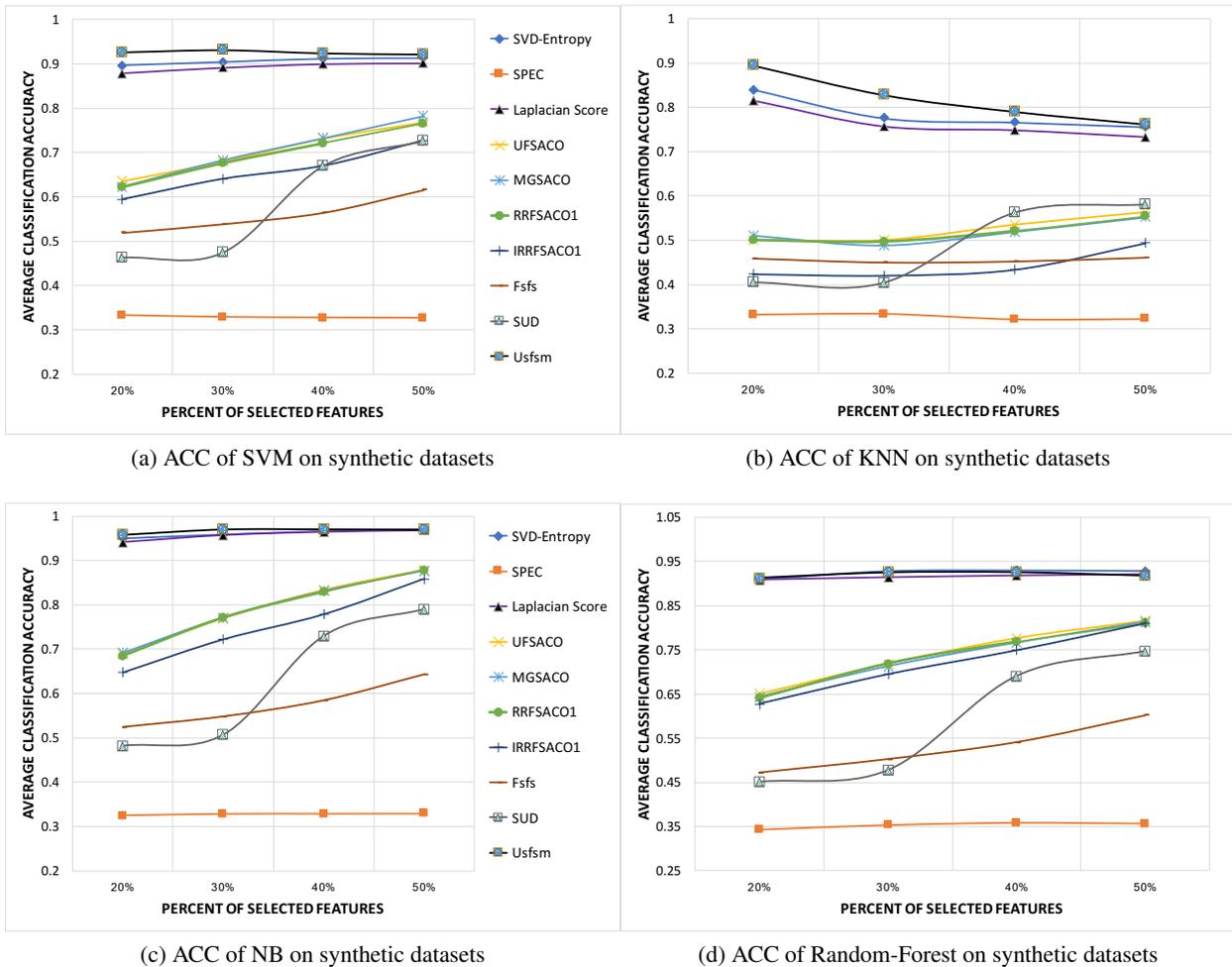


Figure 14: Average classification accuracy of of SVM, KNN ( $K = 3$ ), NB and Random-Forest selecting the 20-50% of features over the synthetic datasets.

our Ph.D. research proposal is introduced; including justification and motivation, the problem to be solved, research questions, hypothesis, our research objectives, and the methodology that will guide our research.

Following the methodology above mentioned, as preliminary result, we have developed a new method, USFSM, for selecting features in unsupervised mixed data. In our method, the relevance of features is measured taking into account the contribution of each feature for defining the cluster structure in the data. The feature evaluation is performed using a kernel and a feature relevance evaluation measure that uses the spectrum of the Normalized Laplacian matrix. After conducting several experiments on different real and synthetic mixed datasets, we can conclude that the proposed method can effectively identify the relevant features in this kind of data. Moreover, our method proved to be better than several unsupervised filter feature selection methods of the state-of-the-art, including SUD, SVD-Entropy, SPEC, Laplacian score, UFSACO, MGSACO, RRFSACO1, IRRFSACO1, and Fsf. The experimental results regarding clustering evaluation measures (ACC and NMI) using the  $k$ -prototypes and EM clustering algorithms show that in most cases USFSM significantly outperforms the other feature selection methods. On the other hand, the results in terms of classification accuracy using four well-known classifiers, namely SVM, KNN, Naive Bayes and Random-Forest show that our method has on average better performance than the other methods. A paper including these results is being prepared for to be submitted to the Journal of Pattern Recognition.

Throughout our preliminary work, we have partially covered the first, second and third points in our proposed methodology. Finally, based on our preliminary results, we conclude that following the proposed methodology our objectives can be reached in the time defined by the Computational Sciences Coordination for accomplishing a Ph.D. degree.

## References

- [1] Isabelle Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [2] Lei Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, 2005.
- [3] Jeffrey L Andrews and Paul D McNicholas. Variable Selection for Clustering and Classification. *Journal of Classification*2, 31(452):136–153, mar 2013.
- [4] H Liu and H Motoda. *Computational methods of feature selection*, volume 198. 2008.
- [5] Gunter Ritter. *Robust Cluster Analysis and Variable Selection*, volume 137. CRC Press, 2015.
- [6] Zheng Alan Zhao and Huan Liu. *Spectral feature selection for data mining*. CRC Press, 2011.
- [7] Sankar K Pal and Pabitra Mitra. *Pattern Recognition Algorithms for Data Mining*. Chapman and {Hall/CRC}, 1 edition, may 2004.
- [8] Igor Kononenko. Estimating attributes: analysis and extensions of RELIEF. In *Machine Learning: ECML-94*, pages 171–182. Springer, 1994.
- [9] Mark A Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato Hamilton, 1999.

- [10] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.
- [11] C Ding and H Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational . . .*, 03(02):185–205, 2005.
- [12] Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu, and Zhihai Wang. A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, 33(1):1–5, 2007.
- [13] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.
- [14] Ronen Feldman and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.
- [15] E B Fowlkes, Ram Gnanadesikan, and John R Kettenring. Variable selection in clustering. *Journal of classification*, 5(2):205–228, 1988.
- [16] J G Dy and C E Brodley. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.
- [17] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature Selection: A Data Perspective. *Journal of Machine Learning Research*, pages 1–73, 2016.
- [18] S Nijjima and Y Okuno. Laplacian linear discriminant analysis approach to unsupervised feature selection. *IEEE/ACM Trans Comput Biol Bioinform*, 6(4):605–614, 2009.
- [19] D. Devakumari and K. Thangavel. Unsupervised adaptive floating search feature selection based on Contribution Entropy. In *2010 International Conference on Communication and Computational Intelligence (INCOCCI)*, pages 623–627. IEEE, 2010.
- [20] Behrouz Zamani Dadaneh, Hossein Yeganeh Markid, and Ali Zakerolhosseini. Unsupervised probabilistic feature selection using ant colony optimization. *Expert Systems with Applications*, 53:27–42, 2016.
- [21] W Rui, J Liu, and Y Jia. Unsupervised feature selection for text classification via word embedding. In *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, pages 1–5, 2016.
- [22] Xiaodong Wang, Xu Zhang, Zhiqiang Zeng, Qun Wu, and Jian Zhang. Unsupervised spectral feature selection with l1-norm graph. *Neurocomputing*, 200:47–54, 2016.
- [23] Elham Hoseini and Eghbal G Mansoori. Selecting discriminative features in social media data: An unsupervised approach. *Neurocomputing*, 205:463–471, 2016.
- [24] Yugen Yi, Wei Zhou, Yuanlong Cao, Qinghua Liu, and Jianzhong Wang. Unsupervised Feature Selection with Graph Regularized Nonnegative Self-representation. In Zhisheng You, Jie Zhou, Yunhong

- Wang, Zhenan Sun, Shiguang Shan, Weishi Zheng, Jianjiang Feng, and Qijun Zhao, editors, *Biometric Recognition: 11th Chinese Conference, CCBR 2016, Chengdu, China, October 14-16, 2016, Proceedings*, pages 591–599. Springer International Publishing, Cham, 2016.
- [25] K Umamaheswari and M Dhivya. D-MBPSO: An Unsupervised Feature Selection Algorithm Based on PSO. In Václav Snášel, Ajith Abraham, Pavel Krömer, Millie Pant, and Azah Kamilah Muda, editors, *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 6th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2015) held in Kochi, India during December 16-18, 2015*, pages 359–369. Springer International Publishing, Cham, 2016.
- [26] Jie Feng, Licheng Jiao, Fang Liu, Tao Sun, and Xiangrong Zhang. Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images. *Pattern Recognition*, 51:295–309, mar 2016.
- [27] Liang Du and Yi-Dong Shen. Unsupervised Feature Selection with Adaptive Structure Learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 209–218. ACM, 2015.
- [28] M A Ambusaidi, X He, and P Nanda. Unsupervised Feature Selection Method for Intrusion Detection System. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*, volume 1, pages 295–301, 2015.
- [29] B Chandra. Chapter 3 - Gene Selection Methods for Microarray Data. In Dhiya Al-JumeilyAbir HussainConor MallucciCarol Oliver, editor, *Applied Computing in Medicine and Health*, Emerging Topics in Computer Science and Applied Computing, pages 45–78. Morgan Kaufmann, Boston, 2016.
- [30] Fumiaki Katagiri and Jane Glazebrook. Overview of mRNA expression profiling using DNA microarrays. *Current Protocols in Molecular Biology*, Chapter 22(SUPPL. 85):Unit 22 4, 2009.
- [31] Kusum Kumari Bharti and Pramod kumar Singh. 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), December 28-30, 2012*, pages 1545–1559. Springer, 2014.
- [32] George Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, 3:1289–1305, 2003.
- [33] Stewart C Bushong and Geoffrey Clarke. *Magnetic resonance imaging: physical and biological principles*. Elsevier Health Sciences, 2013.
- [34] Rafael C Gonzalez and Richard E Woods. Digital image processing. 2002.
- [35] Daniel L Swets and John J Weng. Efficient content-based image retrieval using automatic feature selection. In *Computer Vision, 1995. Proceedings., International Symposium on*, pages 85–90. IEEE, 1995.
- [36] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.
- [37] Wenke Lee, Salvatore J Stolfo, and Kui W Mok. Adaptive intrusion detection: A data mining approach. *Artificial Intelligence Review*, 14(6):533–567, 2000.

- [38] Alexander R De Leon and Keumhee Carrière Chough. *Analysis of mixed data: methods & applications*. CRC Press, 2013.
- [39] Charu C Aggarwal. Outlier analysis. In *Data Mining*, pages 203–263. Springer, 2015.
- [40] Michael J Daniels and T Normand Sharon-lise. Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*, 7(1):1–15, 2006.
- [41] Marc Aerts, Geert Molenberghs, Louise M Ryan, and Helena Geys. *Topics in modelling of clustered data*. CRC Press, 2002.
- [42] Jérôme Paul, Pierre Dupont, and Others. Kernel methods for heterogeneous feature selection. *Neurocomputing*, 169:187–195, 2015.
- [43] Haitao Liu, Ruxiang Wei, and Guoping Jiang. A hybrid feature selection scheme for mixed attributes data. *Computational and Applied Mathematics*, 32(1):145–161, 2013.
- [44] Zenon Gniazdowski and Michał Grabowski. Numerical Coding of Nominal Data. *Zeszyty Naukowe WWSI*, 9(12):53–61, 2015.
- [45] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.
- [46] J Bruin. `newtest`: command to compute new test {@ONLINE}. [http://www.ats.ucla.edu/stat/r/library/contrast\\_coding.htm](http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm), 2011.
- [47] Gauthier Doquire and Michel Verleysen. An Hybrid Approach To Feature Selection for Mixed Categorical and Continuous Data. *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 394–401, 2011.
- [48] J Ruiz-Shulcloper. Pattern recognition with mixed and incomplete data. *Pattern Recognition and Image Analysis*, 18(4):563–576, 2008.
- [49] Rajashree Dash, Rajib Lochan Paramguru, and Rasmita Dash. Comparative analysis of supervised and unsupervised discretization techniques. *International Journal of Advances in Science and Technology*, 2(3):29–37, 2011.
- [50] Erick Cantú-Paz. Supervised and unsupervised discretization methods for evolutionary algorithms. In *Workshop Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 213–216, 2001.
- [51] Alexander Hartemink and D K Gifford. *Principled computational methods for the validation and discovery of genetic regulatory networks*. Massachusetts Institute of Technology. PhD thesis, Ph. D. dissertation, 2001.
- [52] Gavin Brown. A New Perspective for Information Theoretic Feature Selection. In *AISTATS*, pages 49–56, 2009.
- [53] Lei Yu and Huan Liu. Redundancy based feature selection for microarray data. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, volume 5, page 737, New York, New York, USA, 2004. ACM Press.

- [54] Sina Tabakhi and Parham Moradi. Relevance-redundancy feature selection based on ant colony optimization. *Pattern Recognition*, 48(9):2798–2811, 2015.
- [55] Salem Alelyani, Jiliang Tang, and Huan Liu. Feature Selection for Clustering: A Review., 2013.
- [56] Charu C Aggarwal and Chandan K Reddy. *Data clustering: algorithms and applications*. CRC Press, 2013.
- [57] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.
- [58] A K Jain, M N Murty, and P J Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.
- [59] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.
- [60] J B MacQueen. Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.
- [61] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
- [62] Alex Foss, Marianthi Markatou, Bonnie Ray, and Aliza Heching. A semiparametric method for clustering mixed data. *Machine Learning*, 105(3):419–458, 2016.
- [63] Lynette Hunt and Murray Jorgensen. Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361, 2011.
- [64] Ryan P Browne and Paul D McNicholas. Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, 142(11):2976–2984, 2012.
- [65] Brian S Everitt. A finite mixture model for the clustering of mixed-mode data. *Statistics & probability letters*, 6(5):305–309, 1988.
- [66] C J Lawrence and W J Krzanowski. Mixture separation for mixed-mode data. *Statistics and Computing*, 6(1):85–92, 1996.
- [67] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian Score for Feature Selection. In *Advances in Neural Information Processing Systems 18*, volume 186, pages 507–514, 2005.
- [68] Thomas M Cover and Joy A Thomas. *Elements of information theory 2nd edition*. Wiley-interscience, 2006.
- [69] M. Dash, H. Liu, and J. Yao. Dimensionality reduction of unsupervised data. In *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, pages 532–539. IEEE Comput. Soc, 1997.

- [70] M Dash, K Choi, P Scheuermann, and Huan Liu Huan Liu. Feature selection for clustering - a filter solution. *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 115–122, 2002.
- [71] Roy Varshavsky, Assaf Gottlieb, Michal Linial, and David Horn. Novel Unsupervised Feature Filtering of Biological Data. *Bioinformatics*, 22(14):e507 —e513, 2006.
- [72] O Alter and O Alter. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97(18):10101–10106, 2000.
- [73] Monami Banerjee and Nikhil R. Pal. Feature selection with SVD entropy: Some modification and extension. *Information Sciences*, 264:118–134, 2014.
- [74] Fan R K Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.
- [75] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007.
- [76] Rongye Liu, Ning Yang, Xiangqian Ding, and Lintao Ma. An Unsupervised Feature Selection Algorithm: Laplacian Score Combined with {Distance-Based} Entropy Measure. *Intelligent Information Technology Applications, 2007 Workshop on*, 3:65–68, 2009.
- [77] Praisan Padungweang, Chidchanok Lursinsap, and Khamron Sunat. Univariate Filter Technique for Unsupervised Feature Selection Using a New Laplacian Score Based Local Nearest Neighbors. In *Information Processing, 2009. APCIP 2009. Asia-Pacific Conference on*, volume 2, pages 196–200. IEEE, 2009.
- [78] D Garcia-Garcia and R Santos-Rodriguez. Spectral Clustering and Feature Selection for Microarray Data. In *International Conference on Machine Learning and Applications, 2009. {ICMLA} '09*, pages 425–428. IEEE, 2009.
- [79] Pabitra Mitra, C A Murthy, and Sankar K Pal. Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):301–312, 2002.
- [80] Yun Li, Bao-Liang Lu, and Zhong-Fu Wu. Hierarchical Fuzzy Filter Method for Unsupervised Feature Selection. *J. Intell. Fuzzy Syst.*, 18(2):157–169, 2007.
- [81] Sankar K Pal, Rajat K De, and Jayanta Basak. Unsupervised feature evaluation: a neuro-fuzzy approach. *Neural Networks, IEEE Transactions on*, 11(2):366–376, 2000.
- [82] Gerardo Beni and Jing Wang. Swarm intelligence in cellular robotic systems. In Patrick Dario, Paolo and Sandini, Giulio and Aebischer, editor, *Robots and Biological Systems: Towards a New Bionics?*, pages 703–712. Springer, 1993.
- [83] Marco Dorigo and Luca Maria Gambardella. Ant colony system: a cooperative learning approach to the traveling salesman problem. *Evolutionary Computation, IEEE Transactions on*, 1(1):53–66, 1997.

- [84] Sina Tabakhi, Parham Moradi, and Fardin Akhlaghian. An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence*, 32:112–123, 2014.
- [85] Sina Tabakhi, Ali Najafi, Reza Ranjbar, and Parham Moradi. Gene selection for microarray data classification using a novel ant colony optimization. *Neurocomputing*, 168:1024–1036, 2015.
- [86] Laurent El Ghaoui, Guan-Cheng Li, Viet-An Duong, Vu Pham, Ashok N Srivastava, and Kanishka Bhaduri. Sparse Machine Learning Methods for Understanding Large Text Corpora. In *CIDU*, pages 159–173, 2011.
- [87] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon C K Shiu. Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, 2015.
- [88] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou.  $l_2, 1$ -norm regularized discriminative feature selection for unsupervised learning. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1589. Citeseer, 2011.
- [89] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised Feature Selection Using Nonnegative Spectral Analysis. In *AAAI*, 2012.
- [90] Zechao Li and Jinhui Tang. Unsupervised Feature Selection via Nonnegative Spectral Analysis and Redundancy Control. *IEEE Transactions on Image Processing*, 24(12):5343–5355, dec 2015.
- [91] Chenping Hou, Feiping Nie, Dongyun Yi, and Yi Wu. Feature selection via joint embedding learning and sparse regression. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1324. Citeseer, 2011.
- [92] Chenping Hou, Feiping Nie, Xuelong Li, Dongyun Yi, and Yi Wu. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *Cybernetics, IEEE Transactions on*, 44(6):793–804, 2014.
- [93] Zhao Zheng, Wang Lei, and Liu Huan. Efficient Spectral Feature Selection with Minimum Redundancy. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1–6, 2010.
- [94] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342. ACM, 2010.
- [95] David L Donoho and Yaakov Tsaig. Fast solution of  $l_1$ -norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008.
- [96] Martin H C Law, Mario A T Figueiredo, and Anil K Jain. Simultaneous feature selection and clustering using mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1154–1166, 2004.
- [97] Mihaela Breaban and Henri Luchian. A unifying criterion for unsupervised clustering and feature selection. *Pattern Recognition*, 44(4):854–865, 2011.

- [98] E R Hruschka and T F Covoos. Feature selection for cluster analysis: an approach based on the simplified Silhouette criterion. In *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, volume 1, pages 32–38. IEEE, 2005.
- [99] YongSeog Kim, W Nick Street, and Filippo Menczer. Evolutionary model selection in unsupervised learning. *Intelligent data analysis*, 6(6):531–556, 2002.
- [100] YongSeog Kim, W Nick Street, and Filippo Menczer. Data Mining. In John Wang, editor, *Data mining: opportunities and challenges*, chapter Feature Se, pages 80–105. IGI Global, Hershey, PA, USA, 2003.
- [101] Dipankar Dutta, Paramartha Dutta, and Jaya Sil. Simultaneous feature selection and clustering with mixed features by multi objective genetic algorithm. *International Journal of Hybrid Intelligent Systems*, 11(1):41–54, 2014.
- [102] Zhexue Huang. 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)*, pages 21–34. Singapore, 1997.
- [103] Manoranjan Dash and Huan Liu. Feature Selection for Clustering. pages 110–121, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.
- [104] Yun Li, Bao-Liang Lu, and Zhong-Fu Wu. A Hybrid Method of Unsupervised Feature Selection Based on Ranking. In *18th International Conference on Pattern Recognition {(ICPR'06)}*, pages 687–690, Hong Kong, China, 2006.
- [105] E R Hruschka, E R Hruschka, T F Covoos, and N F F Ebecken. Feature selection for clustering problems: a hybrid algorithm that iterates between k-means and a Bayesian filter. In *Fifth International Conference on Hybrid Intelligent Systems, 2005. {HIS} '05*. IEEE, 2005.
- [106] Cláudia Silvestre, Margarida G M S Cardoso, and Mário Figueiredo. Feature selection for clustering categorical data with an embedded modelling approach. *Expert systems*, 32(3):444–453, 2015.
- [107] Julia Handl and Joshua Knowles. Cluster generators for large high-dimensional data sets with large numbers of clusters, 2005.
- [108] Gauthier Doquire, Michel Verleysen, and Others. Mutual information based feature selection for mixed data. In *19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2011)*, 2011.
- [109] Min Wei, Tommy W.S. Chow, and Rosa H.M. Chan. Heterogeneous feature subset selection using mutual information-based feature transformation. *Neurocomputing*, 168:706–718, 2015.
- [110] Wenyin Tang and K.Z. Mao. Feature selection algorithm for mixed data with both nominal and continuous features. *Pattern Recognition Letters*, 28(5):563–571, 2007.
- [111] Yenny Villuendas-Rey, Milton García-Borroto, Miguel A Medina-Pérez, and José Ruiz-Shulcloper. Simultaneous features and objects selection for Mixed and Incomplete data. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 597–605. Springer, 2006.

- [112] José Ruiz-Shulcloper, Mongi A Abidi, and Others. Logical combinatorial pattern recognition: A Review. 2002.
- [113] F Questier, I Arnaut-Rollier, B Walczak, and D L Massart. Application of rough set theory to feature selection for unsupervised clustering. *Chemometrics and Intelligent Laboratory Systems*, 63(2):155–167, 2002.
- [114] Roman W Swiniarski and Andrzej Skowron. Rough set methods in feature selection and recognition. *Pattern Recognition Letters*, 24(6):833–849, 2003.
- [115] Xiao Zhang, Changlin Mei, Degang Chen, and Jinhai Li. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. *Pattern Recognition*, 56:1–15, 2016.
- [116] Zilin Zeng, Hongjun Zhang, Rui Zhang, and Youliang Zhang. A Mixed Feature Selection Method Considering Interaction. *Mathematical Problems in Engineering*, 2015, 2015.
- [117] Qinghua Hu, Jinfu Liu, and Daren Yu. Mixed feature selection based on granulation and approximation. *Knowledge-Based Systems*, 21(4):294–304, 2008.
- [118] Sarah Coppock and Lawrence Mazlack. Rough sets used in the measurement of similarity of mixed mode data. In *Fuzzy Information Processing Society, 2003. NAFIPS 2003. 22nd International Conference of the North American*, pages 197–201. IEEE, 2003.
- [119] Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael von Rhein, and Jan D Reinhardt. Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:132–152, 2013.
- [120] Xin-Ye Li and Li-jie Guo. Constructing affinity matrix in spectral clustering based on neighbor propagation. *Neurocomputing*, 97:125–130, 2012.
- [121] Anneleen Daemen and Bart De Moor. Development of a kernel function for clinical data. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 5913–5917. IEEE, 2009.
- [122] Chung-Chian Hsu, Chien-Hao Kung, and Others. Knowledge Discovery from Mixed Data by Artificial Neural Network with Unsupervised Learning. In *Active Citizenship by Knowledge Management & Innovation: Proceedings of the Management, Knowledge and Learning International Conference 2013*, pages 1295–1302. ToKnowPress, 2013.
- [123] A R De Leon and K C Carriere. A generalized Mahalanobis distance for mixed data. *Journal of Multivariate Analysis*, 92(1):174–185, 2005.
- [124] D. Randall Wilson and Tony R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.
- [125] Tiago R L dos Santos and Luis E Zárate. Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications*, 42(3):1247–1260, 2015.
- [126] Andrés Eduardo Gutierrez-Rodríguez, J Fco Martínez-Trinidad, Milton García-Borroto, and Jesús Ariel Carrasco-Ochoa. Mining patterns for clustering on numerical datasets using unsupervised decision trees. *Knowledge-Based Systems*, 82:70–79, 2015.

- [127] Hendrik Blockeel, Luc De Raedt, and Jan Ramon. Top-down induction of clustering trees. *arXiv preprint cs/0011032*, 2000.
- [128] Xin Sun, Yanheng Liu, Jin Li, Jianqi Zhu, Xuejie Liu, and Huiling Chen. Using cooperative game theory to optimize the feature selection problem. *Neurocomputing*, 97:86–93, 2012.
- [129] Xin Sun, Yanheng Liu, Jin Li, Jianqi Zhu, Huiling Chen, and Xuejie Liu. Feature evaluation and selection with cooperative game theory. *Pattern Recognition*, 45(8):2992–3002, 2012.
- [130] Shiping Wang, Witold Pedrycz, Qingxin Zhu, and William Zhu. Unsupervised feature selection via maximum projection and minimum redundancy. *Knowledge-Based Systems*, 75:19–29, 2015.
- [131] Jundong Li, Xia Hu, Liang Wu, and Huan Liu. Robust Unsupervised Feature Selection on Networked Data. *Proc.SDM*, pages XX–XX, 2016.
- [132] Lei Shi, Liang Du, and Yi-Dong Shen. Robust Spectral Learning for Unsupervised Feature Selection. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 977–982. IEEE, 2014.
- [133] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- [134] Chris Godsil and Gordon F Royle. *Algebraic graph theory*, volume 207. Springer Science & Business Media, 2013.
- [135] Jérôme Paul, Pierre Dupont, and Others. Kernel methods for mixed feature selection. In *ESANN*. Citeseer, 2014.
- [136] László Lovász and Michael D Plummer. Matching theory, volume 121 of North-Holland Mathematics Studies, 1986.
- [137] John C Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, pages 185–208, 1999.
- [138] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.
- [139] George H John and Pat Langley. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.
- [140] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [141] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.
- [142] M Lichman. {UCI} Machine Learning Repository, 2013.
- [143] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.