Unsupervised Feature Selection by Means of External Validity Indices

Javier Béjar
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
bejar@lsi.upc.edu

Abstract

Feature selection for unsupervised data is a difficult task because a reference partition is not available to evaluate the relevance of the features. Recently, different proposals of methods for consensus clustering have used external validity indices to assess the agreement among partitions obtained by clustering algorithms with different parameter values. Theses indices are independent of the characteristics of the attributes describing the data, the way the partitions are represented or the shape of the clusters. This independence allows to use these measures to assess the similarity of partitions with different subsets of attributes.

As for supervised feature selection, the goal of unsupervised feature selection is to maintain the same patterns of the original data with less information. The hypothesis of this paper is that the clustering of the dataset with all the attributes, even when its quality is not perfect, can be used as the basis of the heuristic exploration the space of subsets of features. The proposal is to use external validation indices as the specific measure used to assess well this information is preserved by a subset of the original attributes.

Different external validation indices have been proposed in the literature. This paper will present experiments using the adjusted Rand, Jaccard and Folkes&Mallow indices. Artificially generated datasets will be used to test the methodology with different experimental conditions such as the number of clusters, cluster spatial separanton and the ratio of irrelevant features. The methodology will also be applied to real datasets chosen from the UCI machine learning datasets repository.

1 Introduction

To reduce the dimensionality of a dataset is an important task in machine learning and data mining. It has two advantages, it reduces the computational cost of processing the dataset and it improves the interpretability of the results. Feature selection is a method for dimensionality reduction that eliminates from a dataset all the attributes that are not relevant to the task to be solved. The main advantage of this methods in front of others, like feature extraction, is that the original attributes are preserved and the obtained model is easier to interpret.

Most of the research on feature selection is related to supervised tasks [13]. More recently methods for unsupervised tasks have been appearing in the literature ([14], [16], [5]). Most of these methods are based on characteristics of the features or characteristics of the model obtained by the clustering algorithm. Our proposal is to use measures that use only external properties of the partitions, so they can be applied independently of the characteristics of the features of the dataset or the representation of the partition.

The next section will review the recent literature on unsupervised feature selection. Section 3 will show the different external validity indices used in the experiments. Section 4 will explain the algorithm used to explore the attribute space to perform the attribute selection. Section 5 will show experiments with different artificial and real datasets and the results obtained. Section 6 will summarize the conclusions and future work.

2 Related work

The problem of feature selection can be defined as follows:

> Having a set of features D and a dataset $\mathcal{X}$ that contains N instantiations of the set of features, obtain a subset of features S ($S \subseteq D$) that maintains the relevant information of the dataset for the learning goal.

For supervised features selection the learning goal is defined by a set of *labels* that classify all the instances in the dataset in groups. For unsupervised feature selection the learning goal has to be related to an hypothetical inherent structure of the data.

The classification of methods for feature selection has been established in the supervised feature selection literature, and can be roughly applied also to unsupervised feature selection. The methods can be summarized as:

- *The filter approach:* Features are assessed individually using the task goal and a relevance value is computed that is used to order the features
- *The wrapper approach:* Features are assessed in subsets, an algorithm that performs the learning task is used for this assessment.
- *The embedded approach:* The selection of the most relevant features is an integrated part of the learning task.

Usually feature selection involves two elements, a criteria to be used to measure the relevance of a feature or a subset of features, and a search strategy.

In the filter approach the most important element is the relevance criteria. In supervised feature selection the relevance criteria is related to the labels of the example in the dataset. In unsupervised feature selection there is a more wide diversity of criteria. As the labels are unknown the criteria has to rely on general characteristics of the data such as global or local structure preservation. The search strategy is the criteria to be used to determine the cutting point in the ordered list of features.

There are several examples of this methods in the literature. In [4] is presented a measure based on entropy that evaluates if a dataset presents a structure according with the distribution of the examples. This measure allows to compare different subsets of attributes without clustering the data and to obtain a rank of the attributes. In [9] consensus clustering is used based on clusters obtained by selecting random subsets of features. The consensus allows to compute a ranking of the features based on the correlation of the features within the clusterings. In [22] a correlation measure is used, iteratively features are ranked and selected assuming that the dataset can be obtained by a linear combination of the features. Some other methods rank features using measures that determine how well the local structure is preserved like the Laplacian score [8, 18] or measures related to spectral clustering [23, 25].

In the wrapper approach the search strategy is the most important element. The problem is to be able to select the best subset of features from an exponential number of combinations. This means that an efficient search strategy is crucial. Supervised feature selection uses as a criteria a supervised learning algorithm and performs a search to obtain the minimum subset with the best accuracy. Unsupervised feature selection has to use a clustering algorithm as a method to discover the structure of the data. The search is performed to look for the set of features that preserves the apparent structure of the data obtained by the clustering algorithm.

Some examples of this approach include selection strategies based on internal validation measures the like silhouette index and forward selection as search algorithm [11] or consensus clustering based on random subspaces as measure of feature relevance and genetic algorithms as search strategy [10]. Several approaches propose to extend the problem as a multicriteria optimization problem to solve simultaneously the feature selection problem and the selection of the number of cluster. In these methods, different variations of genetic algorithms are used as search methodology combined with different internal validation measures like the Davis-Bouldin criteria [12, 15, 17, 2].

In the embedded approach a model is assumed in the data. This model usually includes two goals, to determine the number of clusters that describes the structure of the data, and to determine the set of features that better describes the clusters. These two goals are tightly related, so the method is to obtain both simultaneously.

This approach assumes a model based on the mixture of probability distributions, usually gaussians, and extends the model by introducing as parameters of the distribution the selection of the features using weighting schemes (binary or continuous) and the number of components of the mixture. The algorithm used to fit the model is usually the Expectation-Maximization algorithm combined with a model selection strategy based on measures of feature salience, model complexity or cluster quality [14, 5, 3, 20, 24]

3 External validity indices

Cluster validity indices [7] are used to assess the validity of the partitions obtained by a clustering algorithm by comparing its characteristics to a reference or ideal partition. These indices give a relative or absolute value to the similarity among different partitions of the same data.

These indices can be classified in *external criteria*, *internal criteria* and *relative criteria*. The first ones assume that the evaluation of a partition is based on a pre-specified structure that reflects the natural structure of the dataset. For the second ones, the evaluation is performed in terms of quantities obtained from the values of the attributes that describe the clusters, that measure how separated and compact are. The last ones assume that the evaluation is performed comparing the partition to other partitions obtained using different parameters of the clustering algorithm or from other clustering algorithms. In our approach, we are interested on the first kind of validity indices.

External validity indices allow to compare different partitions independently of the type of the attributes used to describe the dataset, the distance function used to measure the similarity of the examples or the model used by the clustering algorithm. In order to do that, only the cluster co-association of the examples between pairs of is used.

The main idea of these indices is that when two examples belong to a natural cluster in the data, they usually appear in the same group when different partitions are obtained. This means that two partitions that maintain these co-association are more similar and represent a similar structure for the dataset.

In order to compute these indices the coincidence of each pair of examples in the groups of two clusterings has to be counted, there are four possible cases:

- The two examples belong to the same class in both partitions (a)
- The two examples belong to the same class in C, but not in P (b)
- The two examples belong to the same class in P, but not in C (c)
- The two examples belong to different classes in both partitions (d)

From this values different similarity indices can be defined for comparing two partitions:

- **Adjusted Rand statistic:**

$$AR = \frac{a - \frac{(a+c)(a+b)}{a+b+c+d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(b+c)}{a+b+c+d}}$$

- **Jaccard Coefficient:**

$$J = \frac{a}{(a+b+c)}$$

- **Folkes and Mallow index:**

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

One of the interesting properties of all this three indices is that they are in the range $[0, 1]$ making more easier to compare their results.

4 Exploration of the attribute space

Being our task unsupervised, there is no ground truth we can use for comparison purposes. To solve this problem, we are going to choose the partition obtained with all the attributes. This partition will be called the *reference partition*. The actual labels of the examples are not needed, if we can define some similarity/quality measure with this partition. The search for a clustering with a reduced number of attributes will be guided by this similarity measure as heuristic.

The task will be to find the subset of attributes that obtains the partition more similar to the reference partition. We will use a wrapper approach. As this task would need to explore all possible subsets we are going to use a best first strategy to reduce the computational cost.

This exploration of the feature space is going to be performed using the *forward selection* strategy. This strategy is also used in wrapper based methods in supervised feature selection [13]. In this case we are going to substitute the supervised algorithm used for the evaluation of attribute subsets by a validity index. This validity index will compute the similarity between the partition obtained with the subset of features and the reference partition.

The forward selection strategy begins with an empty set of attributes and the reference partition. Each step evaluates all partitions that can be obtained by adding one new attribute from the attribute list. If the subset more similar to the reference partition increases the similarity with respect to the one obtained in the previous previous step, that subset is retained, growing the selected list of attributes. The process is iterated using this new subset of attributes. The exploration is stopped when no more attributes can be added obtaining an increase of similarity larger than a threshold α. The detailed algorithm is presented as algorithm 1.

The parameter α will be in the range $[0, 1]$ accordingly with the values of the external index used to measure similarity. There is not a priori good value for this parameter. If the value is closer to 0 less increase of similarity respect to the previous iteration will be needed to include new attributes. If the value is too high the algorithm will stop with only a few attributes selected. As similarity will increase monotonically with the number of attributes a criteria to decide its value is to consider what is the minimum percentage of similarity a new attribute has to add to be considered relevant.

5 Experiments

For the experiments, different datasets will be used. First, synthetic datasets generated with different characteristics will be used, so the feature selection methodology can be tested in a wide range of types of datasets. Specifically, it will be tested the performance with different number of partitions, different spatial separation among the clusters, different number of relevant attributes and different ratio of irrelevant attributes.

The methodology will be also tested with datasets from the UCI machine learning repository [6].

The cluster algorithm used as wrapper algorithm will be k-means using the actual number of clusters as the value of k. To reduce the sensitivity of this algorithm to initialization the k-means++ ([1]) algorithm is used.

5.1 Synthetic datasets

The first set of experiments has been performed using synthetic datasets with different number of clusters and irrelevant attributes. The datasets have been generated using the R package `clusterGeneration` that implements the methodology described in [19].

Algorithm 1 Greedy forward selection

```
Procedure: Greedy_Forward_Selection(DataSet,AttributeSet)
ReferenceClustering= Cluster(DataSet,AttributeSet)
QualityIncrease=true
BestSimilarity=0
BestClustering=∅
CurrentAttributeSet=∅
while QualityIncrease do
    BestNewSimilarity= 0
    BestAttibute= ∅
    BesNewClustering=∅
    for attribute ∈ AttributeSet do
        IAtrributeSet= CurrentAttributeSet + attibute
        CurrenClustering= Cluster(DataSet,IAttributeSet)
        CurrentSimilarity= ValidityIndex(ReferenceClustering, CurrentClustering)
        if CurrentSimilarity>BestNewSimilarity then
            BestNewSimilarity= CurrentSimilarity
            BestNewAttibute= attribute
            BestNewClustering= CurrentClustering
        end
    end
    if BestNewSimilarity-BestSimilarity > α then
        AttributeSet= AttributeSet - BestNewAttibute
        CurrentAttributeSet = CurrentAttributeSet + BestNewAttibute
        BestClustering= BestNewClustering
        BestSimilarity= BestNewSimilarity
    else
        QualityIncrease = false
    end
end
return CurrentAttributeSet
```

The data generated corresponds to sets of ellipsoidal clusters described by attributes modeled using gaussian distributions. These attributes can be divided into relevant and irrelevant attributes. The relevant attributes are the same for all clusters and are generated independently using random orthonormal matrices. Also random rotations are applied to the generated clusters. This means that each cluster has ellipsoidal shape with axis oriented in arbitrary directions. The irrelevant attributes are randomly generated using gaussian distributions with a variance similar to the sample variance of the relevant attributes.

The methodology for generating the data allows to control the minimum degree of separation for a cluster respect to the other clusters. This makes possible to generate datasets with different degrees of separation, from highly overlapped to well separated clusters. Figure 1 shows two datasets using 3 relevant attributes with 3 clusters and different separation.

In the experiments we will vary the number of clusters, the number of attributes, the ratio between the number of irrelevant and irrelevant attributes and the degree of separation between pairs of clusters.

Specifically, the datasets have been generated for 3, 5 and 10 clusters with around 100 examples per cluster. The datasets with 3 and 5 clusters have been generated using 5 and 10 relevant attributes, and the datasets with 10 clusters have been generated with 10 and 20 relevant attributes. For each dataset, a set of irrelevant attributes has been added in the same amount, twice and four times the number of relevant attributes. Clusters are also generated with different degrees of separation, using 0, 0.01 and 0.1 as the values of the parameter that controls the degree of separation in the dataset

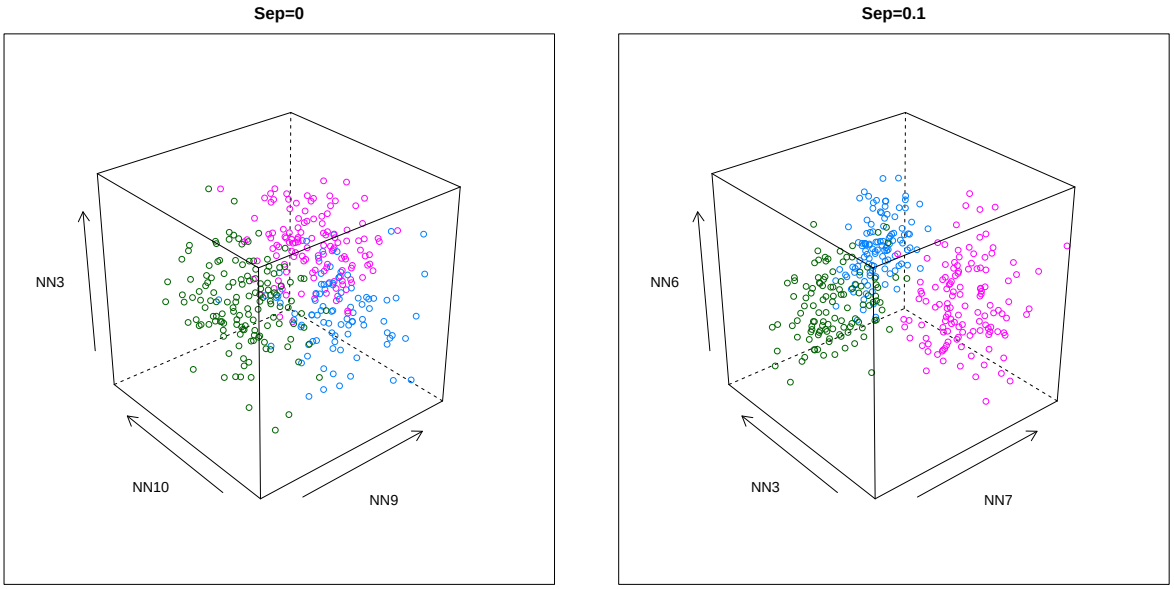

Fig. 1: Synthetic datasets with 3 clusters and separation 0 and 0.1

generation program.

The number of relevant attributes used for generating each dataset is not always the exact number of attributes needed for uncovering the clusters. This number actually depends on the degree of separation among the clusters. To be able to perform fair comparisons the actual list of the relevant attributes for each dataset has been computed using wrapper methodology by performing forward selection using a naive bayes classifier. The mean accuracy of the naive bayes classifiers for the datasets with only the selected attributes is 0.93 with a standard deviation of 0.03.

As performance measures three criteria will be used:

- The similarities to the true labels of the reference partition and the partition obtained using the relevant attributes.
- The number of attributes used to generate the clusters compared to the number attributes after the unsupervised selection.
- The intersection between the list of relevant attributes selected by the unsupervised algorithm and the list of relevant attributes selected by the supervised algorithm.

5.1.1 Comparison of the measures

We will first compare the behavior of the three indices in the experiments. All three use the same information for computing the similarity of partitions, so they are expected to yield similar results. In figure 2 the values of all three measures are plotted against each other for all the experiments with 10 clusters. As it can be seen, the behavior of the Adjusted Rand and the Folkes & Mallow indices are clearly linearly correlated. The Jaccard index shows a slightly nonlinear correlation with the other two indices. This means that the first two indices tend to move their values to the extremes of the interval.

From the distribution of the measures for all the experiments (see figure 3), it can be seen that the distributions of the indices are very close. The Adjusted Rand index is closer to the Jaccard index when there are less clusters and to the Folkes & Mallow index when the number of clusters is larger. A Pearson correlation test over the variables for each different number of clusters shows that the correlation among the three indices is around 0.97 with 99% of confidence. Due to this high correlation among the indices and because the Adjusted Rand index is the most well-known and studied [21], only its value will be shown for the rest of the experiments.

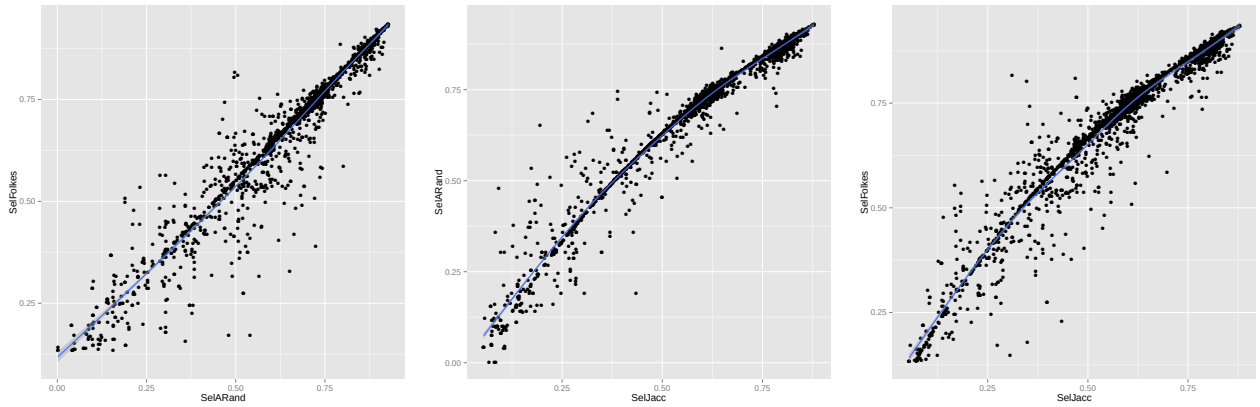

Fig. 2: Scatter plot of all three indices for 10 clusters.

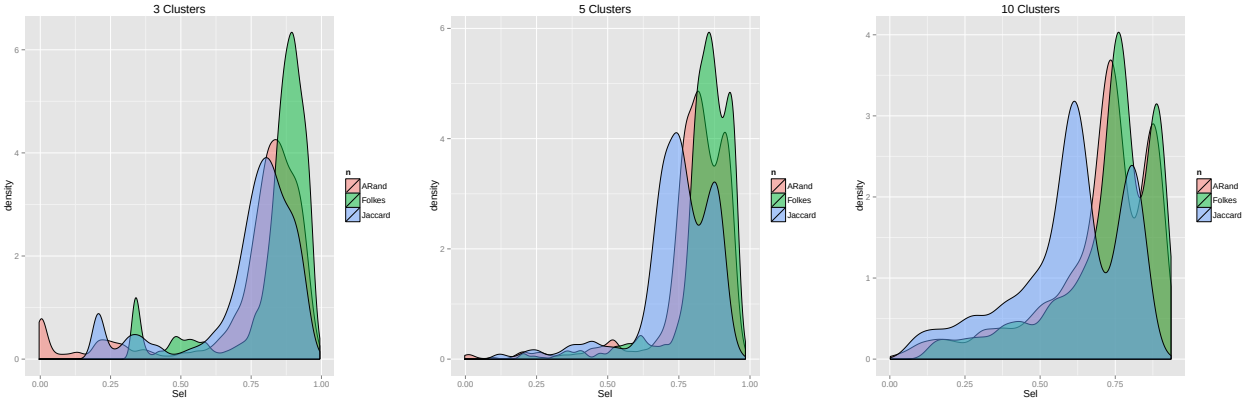

Fig. 3: Distribution of the measures for 3, 5 and 10 clusters.

5.1.2 Cluster separation

Intuition says that the more separated the clusters are, the more easy is for a cluster algorithm to discover the true partition. This means that the reference partition will be more informative and the selection of the attributes will be better.

Figure 4 compares the similarity to the true labels of the partition with only the selected attributes, with the similarity of the partition obtained with all the attributes. Each graphic compares clusters with 3, 5 and 10 partitions with different number of attributes and different ratio of irrelevant attributes. Separation values are 0, 0.01 and 0.1.

Results show that the larger is the cluster separation, the closer is the partition obtained with only the selected attributes to the true partition. This improvement is obtained even when the partition with all the attributes is not so good. It also can be observed a tendency to obtain a better partition after the selection even when clusters are highly overlapped. For similarities under 0.75 a large number of clusterings are largely improved beyond the similarity of the partition with all attributes. This means that a very good reference partition is not strictly necessary for selecting attributes.

The distribution of the similarity of the partitions to the true partition with and without attribute selection also is affected by cluster separation. In figure 5 is represented the distribution of similarities to the true partition for the three values of cluster separation. The graphic shows that the similarity to the true partition is higher when attributes are selected. This difference of distribution is reduced with the cluster separation.

Finally tables 1 and 2 show the number of selected attributes and the irrelevant attributes included in the set of selected attibutes for different number of clusters, number of attributes, ratio of irrelevant attributes and cluster separation. These tables show the mean, the standard deviation and the maximum value.

As it can be seen, the mean number of selected attributes is around the number of attributes

Fig. 4: Similarity of selected and original partition with different levels of separation for 3, 5 and 10 clusters.

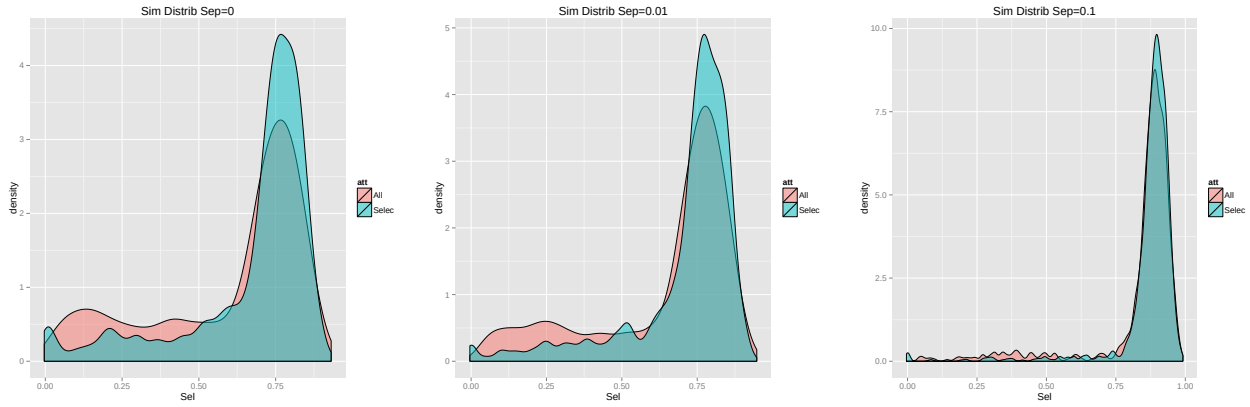

Fig. 5: Distribution of similarity to true partition depending on cluster separation (selected vs all).

used to generate the clusters and sometimes less, depending on the separation of the clusters. The separation affects specially to the number of irrelevant attributes selected. When clusters overlap there is a chance that some irrelevant attributes are also selected.

5.1.3 Number of Clusters

The number of partitions seems to have less influence than the cluster separation on the quality of the final partition after attribute selection. Figure 6 shows the similarity between partitions with 3, 5 and 10 clusters and different levels of separation.

There is a tendency to obtain less improvement in the quality of the final clustering when the number of partitions increases. It is possible to better uncover the clusters when there are fewer of them. It also can be seen that the level of separation of clusters has an influence in the improvement in similarity to the true partition for different number of clusters. More separated clusters allow for less improvement because the reference partition is closer to the true partition.

From table 1 it can be seen that the distribution of the number of selected attributes is independent of the number of clusters, having other factors more influence. Table 2 shows no pattern in the distribution of the number of irrelevant attributes with respect to the number of clusters.

5.1.4 Number of attributes and ratio of irrelevant attibutes

The number of attributes that generates the true partition also have small influence on the final partition. Figure 7 shows the similarity for partitions with 5 attributes with 3 and 5 clusters and with 10 attributes with 3, 5 and 10 clusters. It can be seen that for the same number of generating attributes, the distributions of the similarities are almost the same despite the different number of

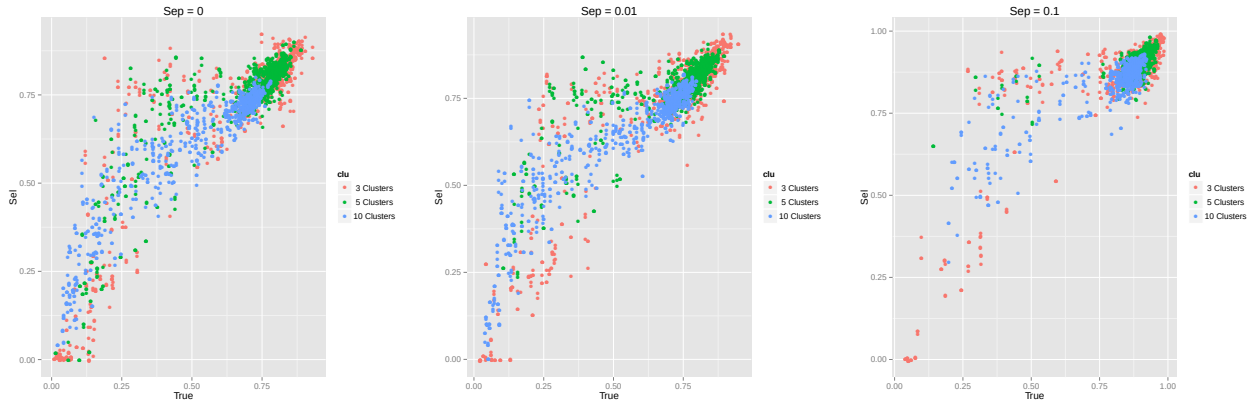

Fig. 6: Similarity of selected and original partition with different levels of separation for 3, 5 and 10 clusters.

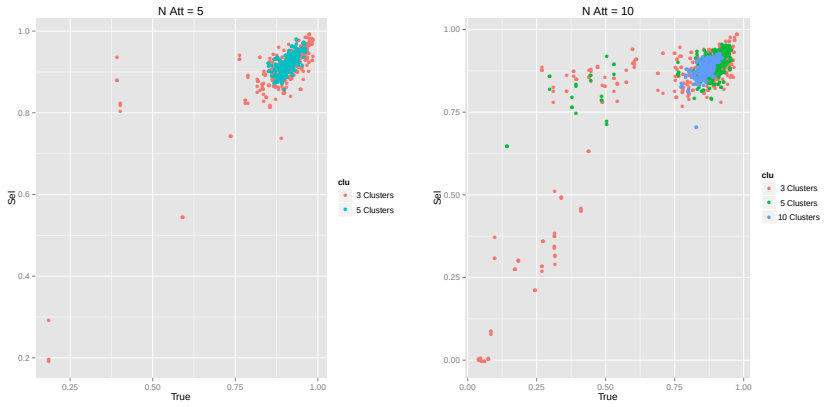

Fig. 7: Similarity of selected and original partition 5 and 10 attributes for 3, 5 and 10 clusters.

clusters. The differences in the means of the distributions are not statistically significative.

Table 1 shows that the number of selected attributes is more related to the number of clusters than the number of generating attributes. It seems that the clusters in the generated datasets can be uncovered by projecting the data in a space of dimensionality around the number of clusters and sometimes even less, as for example the datasets with 10 clusters. Results from table 2 show also that the number of generating attributes is irrelevant, as the distributions are close to each other for most of the experiments.

The ratio of irrelevant attributes has more influence on the similarity of the partition with selected features. Even with well separated clusters the similarity is reduced as the ratio of irrelevant features is increased. This effect increases when the number of generating attributes is also higher. Figure 8 shows the similarity for partitions with 5 and 10 attributes with 1 time, 2 times and 4 times irrelevant attributes for 5 well separated clusters. Figure 9 shows the similarity for partitions with 10 and 20 attributes with 1 time, 2 times and 4 times irrelevant attributes for 10 well separated clusters. It is worth noticing in these two figures that, when the number of attributes is high and there is four times irrelevant attributes, the quality of the partition with selected attributes is systematically better than the reference partition, even when this is very bad.

Table 1 shows that the number of selected attributes does not increases with the increase of irrelevant attributes, but 2 shows that the mean of irrelevant attributes included in the final selection is increased with the number of irrelevant attributes. This means that when the number of irrelevant attributes is large, during feature selection the gain of similarity with the reference partition obtained by the irrelevant attributes, overshadows the gain obtained by the relevant ones. This can be explained in terms of the quality of the reference partition. The actual clusters are very difficult to uncover when the attributes that do not contribute to the patterns are overwhelmingly larger than the ones that

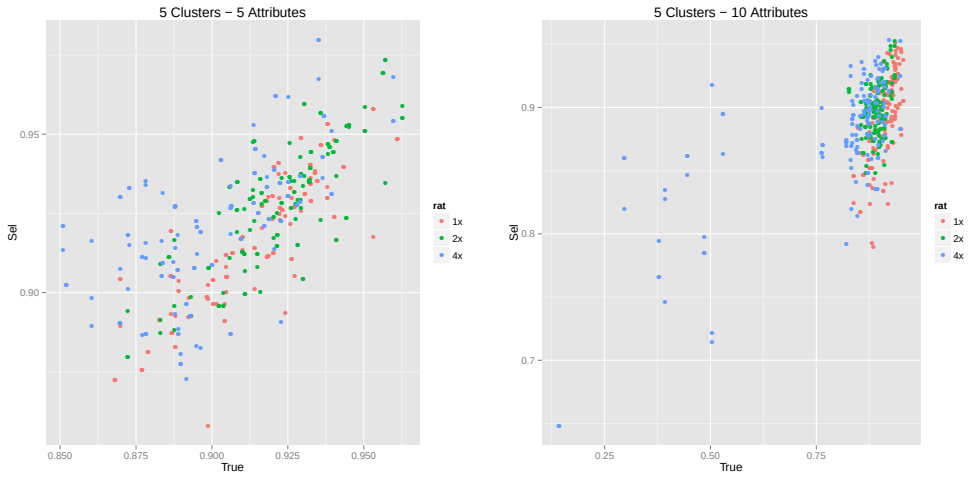

Fig. 8: Similarity of selected and original partition 5 and 10 attributes for 5 clusters and ratio 1, 2 and 4.

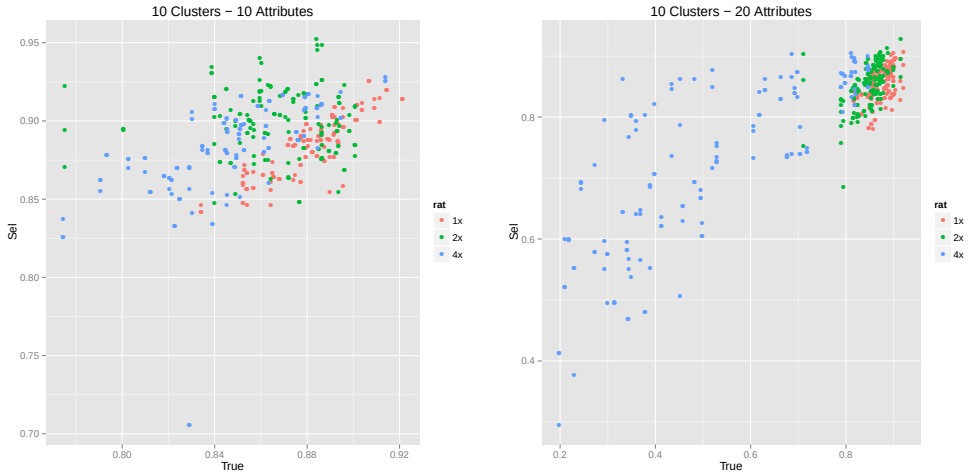

Fig. 9: Similarity of selected and original partition 10 and 20 attributes for 10 clusters and ratio 1, 2 and 4.

define the clusters. A large part of the members of the clusters have been randomly associated, favoring to obtain partitions that deviate from the actual clusters.

5.1.5 Similarity threshold

The parameters that controls the decision of including new attributes to the selected set has more influence when the reference partition is not so good. In figure 10, it can be seen that a difference among the values used for the threshold only appears when the quality of the reference partition is under 0.75. For lower values of the threshold (.01) the quality of the final partition is better. The obvious reason is that the contribution of the relevant attributes to the similarity to the reference partition reduces when the quality of the reference partition is low, so a lower threshold is needed.

5.1.6 Comparison with supervised feature selection

Now we will compare the results using supervised feature selection with the results obtained by unsupervised feature selection. The method used for supervised feature selection is also a wrapper with forward selection. The criteria for adding a new feature is also to obtain an increase of accuracy larger or equal than a threshold. The same values for the threshold where used.

		Ratio								
		1x			2x			4x		
		Separation			Separation			Separation		
Cl.	At.	0	0.01	0.1	0	0.01	0.1	0	0.01	0.1
3	5	4.14 (1.37) 8	4.16 (1.34) 9	3.36 (0.99) 6	4.28 (1.61) 10	4.01 (1.32) 10	3.44 (1.1) 9	4.3 (0.86) 10	4.38 (1.64) 10	3.43 (1.3) 10
	10	5.21 (1.93) 12	4.92 (1.9) 13	4.38 (1.35) 12	5.47 (2.3) 14	5.52 (2.5) 16	4.04 (1.28) 9	6.56 (4.04) 28	5.42 (2.58) 13	4.54 (1.83) 12
5	5	5.38 (1.52) 10	5.14 (1.2) 10	4.62 (0.65) 7	5.36 (1.43) 10	5.26 (1.21) 10	4.59 (0.67) 7	5.26 (1.68) 10	5.22 (1.29) 10	4.46 (0.76) 8
	10	6.36 (1.68) 11	6.48 (1.79) 12	5.54 (1.4) 11	6.52 (1.76) 13	6.34 (1.79) 13	5.4 (1.23) 10	6.12 (2.48) 15	6.04 (1.9) 14	5.43 (1.37) 10
10	10	9.84 (1.16) 14	9.84 (1.15) 13	9.32 (0.61) 13	9.7 (1.45) 18	9.79 (1.19) 14	9.3 (0.61) 12	8.9 (1.6) 15	8.72 (1.55) 13	9.17 (0.62) 12
	20	10.12 (1.7) 16	10.24 (1.7) 18	9.62 (1.2) 14	9.27 (2.3) 16	9 (1.75) 14	9.67 (1.36) 15	7.08 (1.86) 13	7.3 (2.1) 13	8.5 (1.7) 14

Tab. 1: Mean, standard deviation and maximum number of features selected for different experimental conditions

		Ratio								
		1x			2x			4x		
		Separation			Separation			Separation		
Cl.	At.	0	0.01	0.1	0	0.01	0.1	0	0.01	0.1
3	5	0.53 (0.79) 3	0.51 (0.88) 4	0.15 (0.49) 2	0.78 (1.18) 6	0.65 (1.12) 5	0.24 (0.6) 4	1.4 (1.73) 8	1.26 (1.72) 9	0.4 (1) 8
	10	0.43 (0.81) 4	0.4 (0.86) 4	0.22 (0.79) 6	1.38 (1.94) 9	1.37 (2.03) 11	0.28 (0.76) 4	4.2 (3.8) 20	2.56 (2.67) 11	1.49 (1.95) 10
5	5	0.86 (1.33) 5	0.58 (1.03) 5	0.17 (0.44) 2	0.85 (1.3) 6	0.64 (1.04) 5	0.18 (0.45) 2	1.14 (1.63) 7	0.7 (1.16) 6	0.2 (0.52) 3
	10	0.42 (0.79) 4	0.42 (0.82) 3	0.13 (0.42) 3	0.88 (1.34) 6	0.77 (1.31) 7	0.22 (0.55) 3	1.9 (2.3) 12	1.13 (1.65) 10	0.33 (0.67) 4
10	10	0.5 (0.93) 5	0.52 (0.96) 4	0.08 (0.34) 3	0.64 (1.25) 8	0.58 (0.99) 4	0.12 (0.37) 2	1.62 (1.34) 6	0.83 (1.13) 5	0.14 (0.4) 2
	20	0.28 (0.62) 3	0.2 (0.48) 2	0.06 (0.25) 1	0.66 (1.26) 6	0.42 (0.78) 4	0.1 (0.35) 2	1.95 (1.86) 6	1.99 (2.04) 8	0.3 (0.57) 3

Tab. 2: Mean, standard deviation and maximum number of irrelevant features selected for different experimental conditions

		Ratio								
		1x			2x			4x		
		Separation			Separation			Separation		
Cl.	At.	0	0.01	0.1	0	0.01	0.1	0	0.01	0.1
3	5	2.78 (0.82) 5	2.75 (0.72) 5	2.4 (0.58) 4	2.82 (0.83) 5	2.77 (0.74) 5	2.52 (0.67) 5	2.72 (0.87) 5	2.8 (0.82) 5	2.48 (0.62) 4
	10	3.2 (1.07) 6	3.3 (1.13) 6	3.01 (0.96) 6	3.37 (1.23) 7	3.38 (1.2) 7	2.85 (0.92) 6	3.42 (1.19) 6	3.37 (1.22) 7	2.92 (0.9) 5
5	5	4.22 (0.44) 5	4.16 (0.46) 5	3.97 (0.46) 5	4.22 (0.47) 5	4.18 (0.45) 6	3.98 (0.44) 5	4.18 (0.42) 5	4.2 (0.43) 5	3.87 (0.41) 5
	10	4.7 (1.1) 9	4.65 (1.06) 8	4.31 (0.86) 7	4.78 (0.9) 7	4.7 (1.09) 8	4.17 (0.71) 6	4.66 (1.02) 9	4.8 (0.98) 8	4.38 (0.77) 7
10	10	8.53 (0.86) 10	8.54 (0.83) 10	8.09 (0.83) 10	8.49 (0.8) 10	8.5 (0.86) 10	8.09 (0.84) 10	8.6 (0.79) 10	8.5 (0.78) 10	8.02 (0.85) 10
	20	8.8 (1.2) 12	8.79 (1.32) 12	8.27 (1.14) 11	8.83 (1.31) 13	8.76 (1.31) 14	8.16 (1.05) 11	8.76 (1.29) 11	8.7 (1.23) 12	8.17 (0.99) 10

Tab. 3: Mean, standard deviation and maximum number of relevant features selected using supervised methodology for different experimental conditions

		Ratio								
		1x			2x			4x		
		Separation			Separation			Separation		
Cl.	At.	0	0.01	0.1	0	0.01	0.1	0	0.01	0.1
3	5	2.67 (0.75) 5	2.7 (0.7) 5	2.37 (0.58) 4	2.66 (0.77) 5	2.59 (0.77) 4	2.4 (0.66) 5	2.27 (0.93) 5	2.38 (0.87) 5	2.31 (0.6) 4
	10	2.82 (0.96) 6	2.88 (0.96) 6	2.68 (0.88) 5	2.6 (1.09) 5	2.56 (1.11) 6	2.46 (0.76) 4	1.56 (1.2) 5	1.88 (1.1) 4	1.92 (1.03) 4
5	5	4.2 (0.43) 5	4.14 (0.45) 5	3.97 (0.46) 5	4.19 (0.45) 5	4.14 (0.43) 5	3.97 (0.44) 5	3.84 (0.93) 5	4.18 (0.41) 5	3.93 (0.41) 5
	10	4.46 (0.89) 7	4.44 (0.97) 7	4.2 (0.75) 6	4.31 (0.94) 6	4.24 (1) 8	4.02 (0.66) 6	3.27 (1.14) 6	3.87 (0.87) 6	4.02 (0.81) 6
10	10	8.52 (0.84) 10	8.48 (0.8) 10	8.08 (0.82) 10	8.27 (0.97) 10	8.42 (0.83) 10	8.08 (0.85) 10	7.26 (1.39) 10	7.22 (1.32) 9	7.98 (0.87) 10
	20	8.24 (1.08) 11	8.4 (1.1) 12	8.12 (0.96) 10	7.27 (2.02) 10	7.34 (1.38) 10	8.02 (0.94) 10	4.42 (1.7) 9	4.33 (2.01) 10	6.72 (1.6) 10

Tab. 4: Mean number, standard deviation and maximum number of the number of relevant attributes in the intersection for different experimental conditions

Fig. 10: Similarity of selected and original partition 5 and 10 attributes threshold 0.01, 0.03 and 0.05.

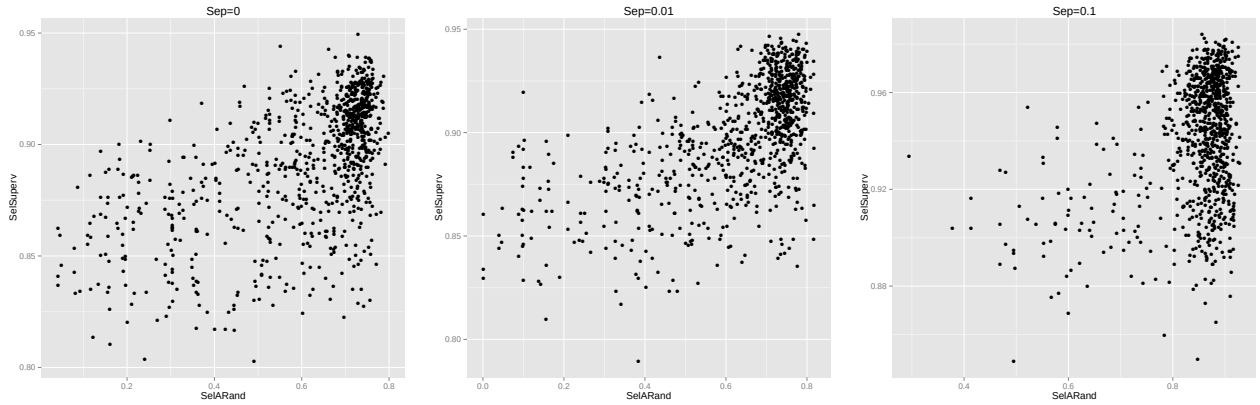

Fig. 11: Adjusted Rand of the unsupervised selection versus accuracy of the supervised selection with 10 clusters for separations 0, 0.01, 0.1.

The learning method used for the wrapper was a naive bayes using gaussian distribution for modeling the attributes. This means that the model is consistent with how the data has been generated. As mentioned before, the mean accuracy obtained is around 0.93.

First we will compare the accuracy obtained in the experiments with the similarity to the true partition measured with the adjusted rand index. This comparison can be done because accuracy is also a measure of partition agreement and the range of the values of both measures are the same.

Figure 11 compares both measures for datasets with 10 clusters and different separations. Results show that separation also affects the accuracy obtained by supervised feature selection, obtaining better accuracy the more separated are the clusters.

Table 3 presents the number of relevant attributes selected using the supervised method. Comparing this result with the results from table 1 it can be observed that, as expected, the number of selected attributes and the variance is lower for the different experimental conditions. For obvious reasons, to have available the ground truth allows to measure more accurately the contribution of each attribute for uncovering of the patterns. But it also has to be noticed that the difference is around one or two extra attributes. This means that even when there is not a good ground truth to compare the selection is still effective.

Table 4 presents the intersection between the sets of selected attributes using supervised and unsupervised information. Is noticeable that when the number of irrelevant attributes is increased one and two times the generating attributes, the agreement among both sets of selected attributes is very high. Only when the number of irrelevant attributes is four times the selected sets differ significantly.

This effect is larger when the clusters are closer. The explanation for this behavior is that when the number of irrelevant attributes is too high it is very difficult to obtain enough ground truth from the dataset. As a consequence, it is more difficult to accurately measure the contribution of each attribute.

5.2 Real datasets

Now we will compare the unsupervised feature selection using real datasets. The datasets are from the UCI machine learning repository [6]. They all are supervised datasets, so the label of the examples is known. Missing values for all the datasets were substituted by the mean or the mode. This comparison is not really fair because for most of the datasets the classes do not correspond to clusters, so it is more difficult for a clustering algorithm to discover exactly the same classes.

For each dataset a supervised algorithm was used to determine the prediction accuracy using all the attributes. A SVM was used for most of the datasets, except for the *wine* dataset, where naive bayes was used and *glass*, where the best accuracy was obtained using 1-knn. These supervised algorithms were used for wrapper forward selection. The number of selected attributes and the prediction accuracy using the selected attributes is shown in table 5. Most of the datasets can be predicted with high accuracy and supervised feature selection can largely reduce the number of attributes. The accuracy with only the selected attributes is only drastically reduced in two datasets.

For the unsupervised feature selection, first clustering was applied to obtain the same number of partitions as supervised classes. The k-means and the gaussian EM algorithms were used to obtain these clusterings. The partitions finally used were the ones the more similar to the supervised classes, according to the adjusted Rand index. The values of similarity are in the first column of table 6. In this table can be also observed that, the unsupervised feature selection method obtains comparable results with the supervised one when the similarity of the clusters and the supervised labels is high. For the datasets with a similarity below 0.3 the method was able to select any of the attributes obtained supervisedly.

As in the experiments with the artificial datasets, when successful, the unsupervised method selects some attributes more than the supervised one, but within the same range. This means that the unsupervised method can successfully discard a large number of irrelevant attributes from the dataset. As can be observed in table 5 the proportion of relevant versus irrelevant attributes is very large for most of the datasets.

To confirm that the unsupervisedly selected features hold actually information from the supervised labels, the same supervised algorithm used before was used to determine the prediction accuracy with only these attributes. The results are shown in the last column of table 6. It can be said that the adjusted Rand similarity is a good predictor of the accuracy. Those datasets with clusters closer to the supervised labels obtain subsets of attributes that represent most of the information from the original labels. This similarity does not have to be very high in order to obtain a good subset of attributes.

In order to obtain a more fair comparison, assuming that unsupervised feature selection would work better when a partition with gaussian clusters is present, the apparent clusters from the datasets have been extracted using gaussian Expectation Maximization. The number of clusters was determined by linear exploration using crossvalidation. For all the datasets, beginning with two clusters, the number was increased until the highest log likelihood was achieved.

Using this number of clusters as classes, the same unsupervised feature selection method was used. Table 7 shows the number of EM clusters, the adjusted Rand index of the supervised labels versus the EM cluster labels, the similarity of the partition with the selected attributes to the EM cluster labels and the number of attributes selected. As expected, the similarity with the supervised labels is largely reduced for most of the datasets. The number of clusters is usually larger than the number of original classes, this is probably the reason of the increased number of selected attributes. The adjusted Rand index of the partition with the selected attributes compared with the EM partition shows that for some of the datasets it is difficult to uncover the seemingly natural clusters. Also, as was seen in the initial experiments, better results are obtained when this similarity is larger than 0.7.

In order to test that the original labels of the datasets can be also predicted using these unsupervisedly selected attributes, the same supervised algorithm used to determine the accuracy with all the

Dataset	Num. Attr.	Num. Classes	Accuracy	Num. Selectes	Accuracy Sel.
Dermatology	35	6	0.97	5	0.86
Glass	10	6	0.72	4	0.57
Ionosphere	35	2	0.95	2	0.91
Iris	4	3	0.98	1	1
Libras	91	15	0.88	5	0.86
Lymph	19	4	0.85	2	0.86
Seeds	8	3	0.98	2	0.95
Soya-small	36	4	0.97	2	1
Vote	17	2	0.96	1	0.95
Wine	14	3	0.98	2	0.94

Tab. 5: Experiments with UCI datasets (Supervised selection)

Dataset	Adj. Rand	Unsup. Sel	Adj. Rand Sel.	Common	Accuracy Sel.
Dermatology	0.31	2	0.02	0	0.41
Glass	0.29	6	0.29	3	0.71
Ionosphere	0.24	3	0.14	0	0.88
Iris	0.9	3	0.9	1	0.96
Libras	0.29	9	0.27	0	0.84
Lymph	0.22	1	0.21	0	0.68
Seeds	0.71	3	0.71	1	0.89
Soya-small	0.93	2	1	1	1.00
Vote	0.68	1	0.83	1	0.95
Wine	0.91	4	0.86	2	0.95

Tab. 6: Experiments with UCI datasets (Unsupervised selection)

Dataset	N. EM Clusters	ARand Sup./EM	Arand Sel./EM	N. Att. Sel.	Accuracy
Dermatology	5	0.6	0.85	4	0.72
Glass	7	0.21	0.46	2	0.41
Ionosphere	10	0.17	0.72	7	0.89
Iris	5	0.51	0.61	1	0.94
Libras	20	0.32	0.71	10	0.81
Lymph	3	0.06	0.54	2	0.75
Seeds	11	0.26	0.55	2	0.88
Soya-small	3	0.65	1.00	1	0.83
Vote	6	0.3	0.53	3	0.81
Wine	4	0.86	0.85	5	0.96

Tab. 7: Experiments with UCI datasets

attributes was applied. The results are shown in the last column of table 7. For most of the datasets the accuracy obtained with the selected attributes is high. This means that most of the structure of the original classes are still in the selected attributes, even when these classes were not used to select them.

6 Conclusions and future work

This paper has presented experiments to study the feasibility of using external validity indices for unsupervised feature selection. The conclusion is that for a limited number of irrelevant features (once or twice the features that generate the clusters) the difference with having the ground truth and performing supervised feature selection is not large. The number of clusters and the number of attributes does not affect largely the effectiveness of the method.

Like for supervised methods, the separability of the patterns has a direct impact on the final selected subset. When clusters are well separated the reference partition used for the selection provides good information that allows to measure more accurately the contribution of the attributes.

As future work, it would be interesting to determine a subset of examples that allow for a better measure of the contribution of the features. This subset of examples would represent the core of the clusters. A possible way for obtaining this core examples would be to use consensus clustering for computing the reference partition. From the consensus it can be determined what examples are more probable to belong to the actual clusters in the data.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein, editors, *SODA*, pages 1027–1035. SIAM, 2007.

[2] Mihaela Breaban and Henri Luchian. A unifying criterion for unsupervised clustering and feature selection. *Pattern Recognition*, 44(4):854–865, 2011.

[3] S. Chang, N. Dasgupta, and L. Carin. A Bayesian Approach to Unsupervised Feature Selection and Density Estimation Using Expectation Propagation. In *CVPR*, pages II: 1043–1050, 2005.

[4] Manoranjan Dash, Kiseok Choi, Peter Scheuermann, and Huan Liu. Feature Selection for Clustering - A Filter Solution. In *ICDM*, pages 115–122. IEEE Computer Society, 2002.

[5] Jennifer G. Dy and Carla E. Brodley. Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research*, 5:845–889, 2004.

[6] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[7] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001. 10.1023/A:1012801612483.

[8] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian Score for Feature Selection. In *NIPS*, 2005.

[9] Y. Hong, S. Kwong, Y. C. Chang, and Q. S. Ren. Consensus unsupervised feature ranking from multiple views. *Pattern Recognition Letters*, 29(5):595–602, April 2008.

[10] Y. Hong, S. Kwong, Y. C. Chang, and Q. S. Ren. Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. *Pattern Recognition*, 41(9):2742–2756, September 2008.

[11] Eduardo R. Hruschka and Thiago F. Covoes. Feature Selection for Cluster Analysis: an Approach Based on the Simplified Silhouette Criterion. In *CIMCA/IAWTIC*, pages 32–38. IEEE Computer Society, 2005.

[12] YongSeog Kim, W. Nick Street, and Filippo Menczer. Evolutionary model selection in unsupervised learning. *Intell. Data Anal*, 6(6):531–556, 2002.

[13] Ron Kohavi and George John. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97:273–324, 1997.

[14] Martin H. C. Law, Anil K. Jain, and Mário A. T. Figueiredo. Feature Selection in Mixture-Based Clustering. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 625–632. MIT Press, 2002.

[15] Ingo Mierswa and Michael Wurst. Information Preserving Multi-Objective Feature Selection for Unsupervised Learning. In Maarten Keijzer et al., editor, *2006 Genetic and Evolutionary Computation Conference (GECCO'2006)*, volume 2, pages 1545–1552, Seattle, Washington, USA, July 2006. ACM Press. ISBN 1-59593-186-4.

[16] Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. Unsupervised Feature Selection Using Feature Similarity. *IEEE Trans. Pattern Anal. Mach. Intell*, 24(3):301–312, 2002.

[17] M. Morita, R. Sabourin, F. Bortolozzi, and C. Y. Suen. Unsupervised Feature Selection Using Multi-Objective Genetic Algorithm for Handwritten Word Recognition. In *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'2003)*, pages 666–670, Edinburgh, Scotland, August 2003.

[18] Satoshi Niijima and Yasushi Okuno. Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection. *IEEE/ACM Trans. Comput. Biology Bioinform*, 6(4):605–614, 2009.

[19] Weiliang Qiu and Harry Joe. Generation of random clusters with specified degree of separation. *Journal of Classification*, 23(2):315–334, 2006.

[20] Adrian E. Raftery and Nema Dean. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473):168–178, March 2006.

[21] D. Steinley. Properties of the hubert-arabie adjusted rand index. *Psychol Methods*, 9 (3):386–396, 2004.

[22] Wei and Billings. Feature Subset Selection and Ranking for Data Dimensionality Reduction. *IEEETPAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 2007.

[23] Lior Wolf and Amnon Shashua. Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weight-Based Approach. *Journal of Machine Learning Research*, 6:1855–1887, 2005.

[24] H. Zeng and Yiu ming Cheung. A new feature selection method for Gaussian mixture clustering. *Pattern Recognition*, 42(2):243–250, February 2009.

[25] Hong Zeng and Yiu ming Cheung. Feature Selection for Clustering on High Dimensional Data. In Tu Bao Ho and Zhi-Hua Zhou, editors, *PRICAI*, volume 5351 of *Lecture Notes in Computer Science*, pages 913–922. Springer, 2008.