Clustering and Selecting Categorical Features

Cláudia Silvestre, Margarida Cardoso, and Mário Figueiredo

Escola Superior de Comunicação Social, Lisboa, Portugal
ISCTE, Business School, Lisbon University Institute, Lisboa, Portugal
Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal
csilvestre@escs.ipl.pt.
margarida.cardoso@iscte.pt.
mario.figueiredo@lx.it.pt

Abstract. In data clustering, the problem of selecting the subset of most relevant features from the data has been an active research topic. Feature selection for clustering is a challenging task due to the absence of class labels for guiding the search for relevant features. Most methods proposed for this goal are focused on numerical data. In this work, we propose an approach for clustering and selecting categorical features simultaneously. We assume that the data originate from a finite mixture of multinomial distributions and implement an integrated expectation-maximization (EM) algorithm that estimates all the parameters of the model and selects the subset of relevant features simultaneously. The results obtained on synthetic data illustrate the performance of the proposed approach. An application to real data, referred to official statistics, shows its usefulness.

Keywords: Cluster analysis, finite mixtures models, EM algorithm, feature selection, categorical variables

1 INTRODUCTION

Feature selection is considered a fundamental task in several areas of application that deal with large data sets containing many features, such as data mining, machine learning, image retrieval, text classification, customer relationship management, and analysis of DNA micro-array data. In these settings, it is often the case that not all the features are useful: some may be redundant, irrelevant, or too noisy. Feature selection extracts valuable information from the data sets, by choosing a meaningful subset of all the features. Some benefits of feature selection include reducing the dimensionality of the feature space, removing noisy features, and providing better understanding of the underlying process that generated the data.

In supervised learning, namely in classification, feature selection is a clearly defined problem, where the search is guided by the available class labels. In contrast, for unsupervised learning, namely in clustering, the lack of class information makes feature selection a less clear problem and a much harder task.

An overview of the methodologies for feature selection as well as guidance on different aspects of this problem can be found in [1], [2] and [3].

In this work, we focus on feature selection for clustering categorical data, using an embedded approach to select the relevant features. We adapt the approach developed by Law et al. [4] for continuous data that simultaneous clusters and selects the relevant subset of features. The method is based on a minimum message length (MML) criterion [5] to guide the selection of the relevant features and an *expectation-maximization* (EM) algorithm [6] to estimate the model parameters. This variant of the EM algorithm seamlessly integrates model estimation and feature selection into a single algorithm. We work within the commonly used framework for clustering categorical data that assumes that the data originate from a multinomial mixture model. We assume that the number of components of the mixture model is known and implement a new EM variant following previous work in [7].

2 RELATED WORK

Feature selection methods aim to select a subset of relevant features from the complete set of available features in order to enhance the clustering analysis performance. Most methods can be categorized into four classes: filters, wrappers, hybrid, and embedded.

The filter approach assesses the relevance of features by considering the intrinsic characteristics of the data and selects a feature subset without resorting to clustering algorithm. Some popular criteria used to evaluate the goodness of a feature or of a feature subset are distance, information, dependency, or consistency measures. Some filter methods produce a feature ranking and use a threshold to select the feature subset. Filters are computationally fast and can be used in unsupervised learning.

Wrapper approaches include the interaction between the feature subset and the clustering algorithm. They select the feature subset, among various candidate subsets of features that are sequentially generated (usually in a forward or backward way), in an attempt to improve the clustering algorithm results. Usually, wrapper methods are more accurate than filters, but even for algorithms with a moderate complexity, the number of iterations that the search process requires results in a high computational cost.

Hybrid methods aim at taking advantage of the best of both worlds (filters and wrappers). The main goal of hybrid approaches is to obtain the efficiency of filters and the accuracy of wrappers. Usually, hybrid algorithms use a filter method to reduce the search space that will subsequently be considered by the a wrapper. Hybrid methods are faster than wrappers, but slower than filters.

In embedded methods, the feature selection is included into the clustering algorithm, thus fully exploiting the interplay between the selected features and the clustering task. Embedded methods are reported to be much faster than wrappers, although their performance also depends on the clustering algorithm [8].

In clustering problems, feature selection is both challenging and important. Filters used for supervised learning can be used for clustering since they do not resort to class labels. The vast majority of work on feature selection for clustering has focused on numerical data, namely on Gaussian-mixture-based methods (e. g. [9], [4], and [10]). In contrast, work on feature selection for clustering categorical data is relatively rare [11].

Finite mixture models are widely used for cluster analysis. These models allow a probabilistic approach to clustering in which model selection issues (e.g., number of clusters or subset of relevant features) can be formally addressed. Some advantages of this approach are: it identifies the clusters, it is able to deal with different types of features measurements, and it outperforms more traditional approaches (e.g., k-means). Finite mixture models assume specific intra-cluster probability functions, which may belong to the same family but differ in the parameter values. The purpose of model estimation is to identify the clusters and estimate the parameters of the distributions underlying the observed data within each cluster. The maximum likelihood estimators cannot be found analytically, and the EM algorithm [6] has been often used as an effective method for approximating the estimates. To our knowledge there is only one proposal [11] within this setting for clustering and selecting categorical features. In his work, Talavera presents a wrapper and a filter to select categorical features. The proposed wrapper method, EM-WFS (EM wrapper with forward search), combines EM with forward feature selection. Assuming that the feature dependencies play a crucial role in determining the feature importance for clustering, a filter ranker based on a mutual information measure, EM-PWDR (EM pairwise dependency ranker), is proposed. In supervised learning, filter approaches usually measure the correlation of each feature with the class label by using distance, information, or dependency measures [12]. Assuming that, in the absence of class labels, we can consider as irrelevant those features that exhibit low dependency with the other features [13]. Under this assumption, the proposed filter considers as good candidates to be selected the highly correlated features with other features. Feature subset evaluation criteria like scatter separability or maximum likelihood seem to be more efficient for the purpose of clustering than the dependence between features. In our work, we propose an embedded method for feature selection, using a minimum message length model selection criterion to select the relevant features and a new EM algorithm for performing model-based clustering.

3 THE MODEL

Let $Y = \left[\underline{y}_1, \ldots, \underline{y}_n\right]'$ be a sample of n independent and identically distributed random variables/features, where $\underline{y} = (Y_1, \ldots, Y_L)$ is a L-dimensional random vector. It is said that $\underline{y}$ follows a K component finite mixture distribution if its

loglikelihood can be written as

$$\log \prod_{i=1}^{n} f(\underline{y}_i|\underline{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k f(\underline{y}_i|\underline{\theta_k})$$

where $\alpha_1, \ldots, \alpha_K$ are the mixing probabilities ($\alpha_k \geq 0, k = 1, .., K$ and $\sum_{k=1}^{K} \alpha_k = 1$), $\underline{\theta} = (\underline{\theta}_1, .., \underline{\theta}_K, \alpha_1, .., \alpha_K)$ the set of all the parameters of the model and $\underline{\theta}_k$ is the set of parameters defining the k-th component. In our case, for categorical data, $f(.)$ is the probability function of a multinomial distribution.

Assuming that the features are conditionally independent given the component-label, the log-likelihood is

$$\log \prod_{i=1}^{n} f\left(\underline{y}_i|\underline{\theta}\right) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} f(\underline{y}_{li}|\underline{\theta}_{lk})$$

The maximum likelihood estimators cannot be found analytically, and the EM algorithm has been often used as an effective method for approximating the corresponding estimates. The basic idea behind the EM algorithm is regarding the data Y as incomplete data, clusters allocation being unknown. In finite mixture models, variables $Y_1, \ldots, Y_L$ (the incomplete data) are augmented by a component-label latent variables $\underline{z} = (Z_1, \ldots, Z_K)$ which is a set of K binary indicator latent variables, that is, $\underline{z_i} = (Z_{1i}, \ldots, Z_{Ki})$, with $Z_{ki} \in \{0, 1\}$ and $Z_{ki} = 1$ if and only if the density of $\underline{y_i} \in C_k$ (component k) implying that the corresponding probability function is $f(\underline{y}_i|\underline{\theta}_k)$. Assuming that the $Z_1, \ldots, Z_k$ are i.i.d., following a multinomial distribution of K categories, with probabilities $\alpha_1, \ldots, \alpha_K$, the log-likelihood of a complete data sample $(\underline{y}, \underline{z})$, is given by

$$\log f(\underline{y}_i, \underline{z}_i|\underline{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ki} \log \left[\alpha_k f(\underline{y}_i|\underline{\theta}_k)\right]$$

The EM algorithm produces a sequence of estimates $\hat{\underline{\theta}}(t)$, $t = 1, 2, \ldots$ until some convergence criterion is met.

3.1 Feature Saliency

The concept of feature saliency is essencial in the context of the feature selection methodology. There are different definitions of feature saliency/(ir)relevancy. Law et al. [4] adopt the following definition: a feature is irrelevant if its distribution is independent of the cluster labels i.e. an irrelevant feature has a common to all clusters probability function.

Lets denote the probability function of relevant and irrelevant features by p(.) and q(.), respectively. For categorical features, p(.) and q(.) refer to multinomial distributions. Let $B_1, \ldots, B_L$ be the binary indicators of the features relevancy, where $B_l = 1$ if the feature l is relevant and zero otherwise.

Using this definition of feature irrelevancy the log-likelihood becomes

$$\log \prod_{i=1}^{n} f\left(\underline{y}_i|\underline{\theta}\right) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} \left[\mathrm{p}\left(\underline{y}_{li}|\underline{\theta}_{lk}\right)\right]^{B_l} \left[\mathrm{q}\left(\underline{y}_{li}|\underline{\theta}_{l}\right)\right]^{1-B_l}$$

Defining *feature saliency* as the probability of the feature being relevant, $\rho_l = P(B_l = 1)$ the log-likelihood is (the proof is in [4]):

$$\log \prod_{i=t}^{n} f\left(\underline{y}_i|\underline{\theta}\right) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} \left[\rho_l \mathrm{p}\left(\underline{y}_{li}|\underline{\theta}_{lk}\right) + (1-\rho_l)\,\mathrm{q}\left(\underline{y}_{li}|\underline{\theta}_{l}\right)\right]$$

The features' saliencies are unknown and they are estimated using an EM variant based on the MML criterion. This criterion encourages the saliencies of the relevant features to go to 1 and of the irrelevant features to go to zero, pruning the features' set.

4 The Proposed Method

We propose an embedded approach for clustering categorical data, assuming hat the data are originate from a multinomial mixture and the number of mixture components is known. The new EM algorithm is implemented using an MML criterion to estimate the mixture parameters, including the features' saliencies. This work extends that of Law et al. [4] dealing with categorical features.

4.1 The Minimum Message Length (MML) Criterion

The MML-type criterion chooses the model providing the shortest description (in an information theory sense) of the observations [5]. According to Shannon's information theory, if Y is some random variable with probability distribution $p(y|\underline{\theta})$, the optimal code-length for an outcome y is $l(y|\underline{\theta}) = \log_2 p(y|\underline{\theta})$, measured in bits and ignoring that $l(y)$ should be integer [14]. When the parameters, $\underline{\theta}$, are unknown they need to be encoded, so the total message length is given by $l(\underline{y}, \underline{\theta}) = l(\underline{y}|\underline{\theta}) + l(\underline{\theta})$, where the first part encodes the observation $\underline{y}$, and the second the parameters of the model.

Under the MML criterion, for categorical features, the estimate of $\underline{\theta}$ is the one that minimizes the following description length function:

$$\mathrm{l}(\underline{y}, \underline{\theta}) = -\log f(\underline{y}|\underline{\theta}) + \frac{K+L}{2}\log n + \sum_{l=1,\rho_l\neq 0}^{L} \frac{c_l - 1}{2} \sum_{k=1}^{K} \log(n\alpha_k\rho_l)$$
$$+ \sum_{l=1,\rho_l\neq 1}^{L} \frac{c_l - 1}{2} \log(n(1-\rho_l))$$

where c_l is the number of categories of feature Y_l.

A Dirichlet-type *prior* (a natural conjugate *prior* of the multinomial) is used for the saliencies,

$$p(\rho_1, \ldots, \rho_L) \propto \prod_{l=1}^{L} \rho_l^{\frac{-Kc_l}{2}} \left(1-\rho_l\right)^{\frac{c_l}{2}}.$$

As a consequence, the MAP (*maximum a posterior*) parameters estimators are obtained when minimizing the proposed description length function, $\mathrm{l}(\underline{y}, \underline{\theta})$.

4.2 The Integrated EM

To estimate all the parameters of the model, we implemented a new version of the EM algorithm integrating clustering and feature selection - the *integrated Expectation-Maximization* (iEM) algorithm. This algorithm complexity is the same as the standard EM for mixture of multinomials. The iEM algorithm to maximize $[-\mathrm{l}(\underline{y}, \underline{\theta})]$ has two steps:

E-step: Compute

$$\mathrm{P}\left[Z_{ki}=1|\underline{Y}_i, \underline{\theta}\right] = \frac{\alpha_k \prod_{l=1}^{L}\left[\rho_l \mathrm{p}\left(\underline{y}_{li}|\underline{\theta}_{lk}\right) + \left(1-\rho_l\right)\mathrm{q}\left(\underline{y}_{li}|\underline{\theta}_{l}\right)\right]}{\sum_{k=1}^{K}\alpha_k \prod_{l=1}^{L}\left[\rho_l \mathrm{p}\left(\underline{y}_{li}|\underline{\theta}_{lk}\right) + \left(1-\rho_l\right)\mathrm{q}\left(\underline{y}_{li}|\underline{\theta}_{l}\right)\right]} \tag{1}$$

M-step: Update the parameter estimates according to

$$\hat{\alpha}_k = \frac{\sum_i \mathrm{P}\left[Z_{ki}=1|\underline{Y}_i, \underline{\theta}\right]}{n}, \tag{2}$$

$$\hat{\theta}_{lkc} = \frac{\sum_i u_{lki} y_{lci}}{\sum_c \sum_i u_{lki} y_{lci}}, \tag{3}$$

$$\hat{\rho}_l = \frac{\max\left(\sum_{ik} u_{lki} - \frac{K(c_l-1)}{2}, 0\right)}{\max\left(\sum_{ik} u_{lki} - \frac{K(c_l-1)}{2}, 0\right) + \max\left(\sum_{ik} v_{lki} - \frac{c_l-1}{2}, 0\right)} \tag{4}$$

where

$$u_{lki} = \frac{\rho_l \mathrm{p}\left(\underline{y}_{li}|\underline{\theta}_{lk}\right)}{\rho_l \mathrm{p}\left(\underline{y}_{li}|\underline{\theta}_{lk}\right) + \left(1-\rho_l\right)\mathrm{q}\left(\underline{y}_{li}|\underline{\theta}_{l}\right)} \mathrm{P}\left[Z_{ki}=1|\underline{Y}_i, \underline{\theta}\right]$$

$$v_{lki} = \mathrm{P}\left[Z_{ki}=1|\underline{Y}_i, \underline{\theta}\right] - u_{lki}$$

After running the iEM, usually the saliencies are not zero or one. Our goal is to reduce the set of initial features, so we check if pruning the feature which

has the smallest saliency produces a lower message length. This procedure is repeated until all the features have their saliencies equal to zero or one. At the end, we will choose the model having the minimum message length value. The proposed algorithm is summarized in Figure 1.

Fig. 1. The iEM algorithm for clustering and selecting categorical features.

Input: data $Y = \left[\underline{y}_1, \ldots, \underline{y}_n\right]'$ where $\underline{y} = (Y_1, \ldots, Y_L)$
the number of components K
mimimum increasing threshold for the likelihood function δ
Ouput: feature saliencies $\{\rho_1, \ldots, \rho_L\}$
mixture parameters $\{\underline{\theta}_{1k}, \ldots, \underline{\theta}_{LK}\}$ and $\{\alpha_1, \ldots, \alpha_K\}$
parameters of common distribution $\{\underline{\theta}_1, \ldots, \underline{\theta}_L\}$

Initialization: initialization of the parameters resorts to the empirical distribution:
set the parameters $\underline{\theta}_{lk}$ of the mixture components
$\mathrm{p}(\underline{y}_l|\underline{\theta}_{lk})$, $(l = 1, \ldots, L\ ;\ k = 1, \ldots, K)$
set the common distribution parameters $\underline{\theta}_l$, to cover all the data
$\mathrm{q}(\underline{y}_l|\underline{\theta}_l)$, $(l = 1, \ldots, L)$
set all features saliencies
$\rho_l = 0.5$ $(l = 1, \ldots, L)$
store the initial log-likelihood
store the initial message length (iml)
$mindl \leftarrow iml$
continue $\leftarrow 1$
while continue **do**
  while increases on log-likelihood are above δ **do**
    M-step according to (2), (3) and (4)
    E-step according to (1)
    if (feature l is relevant) $\rho_l = 1, \mathrm{q}(\underline{y}_l|\underline{\theta}_l)$ is pruned
    if (feature l is irrelevant) $\rho_l = 0, \mathrm{p}(\underline{y}_l|\underline{\theta}_{lk})$ is pruned for all k
    Compute the log-likelihood and the current message length (ml)
  end while
  if $ml < mindl$
    $mindl \leftarrow ml$
    update all the parameters of the model
  end if
  if there are saliencies, $\rho_l \notin \{0, 1\}$
    prune the variable with the smallest saliency
  else
    continue $\leftarrow 0$
  end if
end while
The best solution including the saliencies corresponds to the final $mindl$ obtained.

5 NUMERICAL EXPERIMENTS

For a L-variate multinomial we have

$$f\left(\underline{y}_i|\underline{\theta}\right) = \prod_{l=1}^{L}\left[n!\prod_{c=1}^{c_l}\frac{\theta_{lc}^{y_{lci}}}{(y_{lci})!}\right]$$

where c_l is the number of categories of feature Y_l.

5.1 Synthetic Data

We use two types of synthetic data: in the first type the irrelevant features have exactly the same distribution for all components. Since with real data, the irrelevant features could have little (non relevant) differences between the components, we consider a second type of data where we simulate irrelevant features with similar distributions between the components. In both cases, the irrelevant features are also distributed according to a multinomial distribution. Our approach is tested with 8 simulated data sets. We ran the proposed EM variant (iEM) 10 times and chose the best solution. According to the obtained results using the iEM, the estimated probabilities corresponding to the categorical features almost exactly match the actual (simulated) probabilities. Two of our data sets are presented in tables 1 and 2.

In Table 1 results refer to one data set with 900 observations, 4 categorical features and 3 components with 200, 300 and 400 observations. The first two features are relevant with 2 and 3 categories respectively, the other two are irrelevant and have 3 and 2 categories each. These irrelevant features have the same distribution for the 3 components. In Table 2 the data set has 900 observations and 5 categorical features. The features 1, 4 and 5 have 3 categories each and the features 2 and 3 have 2 categories. The first three features are relevant and the last two are irrelevant, with similar distributions between components.

5.2 Real Data

An application to real data referred to european official statistics (EOS) illustrates the usefulness of the proposed approach. This EOS data set originates from a survey on perceived quality of life in 75 european cities, with 23 quality of life indicators (clustering base features). For modeling purposes the original answers - referring to each city respondents- are summarized into: Scale 1)- *agree* (including strongly agree and somewhat agree) and *disagree* (including somewhat disagree and strongly disagree) and Scale 2)- *satisfied* (including very satisfied and rather satisfied) and *unsatisfied* (including rather unsatisfied and not at all satisfied).

A two-step approach is implemented: firstly, the number of clusters is determined based on MML criterion - see [15]; secondly, the proposed iEM algorithm is applied 10 times and the solution that has the lower *message length* is chosen.

Table 1. iEM results for a synthetic data set where irrelevant features have the same distributions between components.

	Synthetic data			The algorithm's results			
	Component			Component			Saliency
	1 Dim. 200 $\alpha = 0.22$	2 Dim. 300 $\alpha = 0.33$	3 Dim. 400 $\alpha = 0.45$	1 $\alpha = 0.22$	2 $\alpha = 0.33$	3 $\alpha = 0.45$	mean std. dev. (of 10 runs)
Feature 1	0.20	0.70	0.50	0.20	0.70	0.50	$\bar{x} = .99$
relevant	0.80	0.30	0.50	0.80	0.30	0.50	$s = .01$
Feature 2	0.20	0.10	0.60	0.20	0.10	0.60	$\bar{x} = .97$
relevant	0.70	0.30	0.20	0.70	0.30	0.20	$s = .11$
	0.10	0.60	0.20	0.10	0.60	0.20	
Feature 3	0.50			0.50			$\bar{x} = 0$
irrelevant	0.20			0.20			$s = 0$
	0.30			0.30			
Feature 4	0.40			0.40			$\bar{x} = 0.10$
irrelevant	0.60			0.60			$s = 0.11$

Table 2. iEM results for a synthetic data set where irrelevant features have similar distributions between components.

	Synthetic data		The algorithm's results		
	Component		Component		Saliency
	1 Dim. 400 $\alpha = 0.44$	2 Dim. 500 $\alpha = 0.56$	1 $\alpha = 0.44$	2 $\alpha = 0.56$	mean std. dev. (of 10 runs)
Feature 1	0.70	0.10	0.70	0.10	$\bar{x} = 1$
relevant	0.20	0.30	0.20	0.30	$s = 0$
	0.10	0.60	0.10	0.60	
Feature 2	0.20	0.70	0.20	0.69	$\bar{x} = 1$
relevant	0.80	0.30	0.80	0.31	$s = 0$
Feature 3	0.40	0.60	0.40	0.60	$\bar{x} = .99$
relevant	0.60	0.40	0.60	0.40	$s = .01$
Feature 4	0.5	0.49	0.50		$\bar{x} = .04$
irrelevant	0.20	0.22	0.20		$s = .06$
	0.30	0.29	0.30		
Feature 5	0.30	0.31	0.31		$\bar{x} = .04$
irrelevant	0.30	0.30	0.30		$s = .07$
	0.40	0.39	0.39		

Features' saliencies mean and standard deviations over 10 runs are presented in Table 3.

Applying the iEM algorithm to group the 75 European cities into 4 clusters, 2 quality of life indicators are considered irrelevant: *Presence of foreigners is good for the city* and *Foreigner here are well integrated*, meaning that the opinions re-

Table 3. Features' saliencies: mean and standard deviation of 10 runs

	Saliency	
Features	mean	std. dev.
Satisfied with sport facilities	0.95	0.16
Satisfied with beauty of streets	0.99	0.03
City committed to fight against climate change	0.99	0.04
Satisfied with public spaces	0.93	0.18
Noise is a big problem here	0.74	0.25
Feel safe in this city	0.78	0.34
Feel safe in this neighborhood	0.71	0.35
Administrative services help efficiently	0.99	0.02
Satisfied with green space	0.62	0.17
Resources are spent in a responsible way	0.72	0.21
Most people can be trusted	0.74	0.21
Satisfied with health care	0.64	0.11
Poverty is a problem	0.54	0.34
Air pollution is a big problem here	0.65	0.16
It is easy to find a good job here	0.51	0.21
This is a clean city	0.73	0.20
Satisfied with outdoor recreation	0.77	0.23
Easy to find good housing at reasonable price	0.44	0.09
City is healthy to live in	0.54	0.21
Satisfied with cultural facilities	0.5	0.22
Satisfied with public transport	0.28	0.14
Foreigner here are well integrated	0.26	0.24
Presence of foreigners is good for the city	0.38	0.4

garding these features are similar for all the clusters. In fact, most of the citizens (79%) agree that the presence of foreigners is good for the city but they do not agree that foreigners are well integrated (only 39 % agree). Clustering results along with features' saliencies are presented in Table 4 - reported probabilities regard the agree and satisfied categories.

According to the obtained results we conclude that most respondents across all surveyed cities feel safe in their neighborhood and in their city. In cluster 1 cities air pollution and noise are relevant problems and it is not easy to find good housing at reasonable price. It is not easy to find a job in cities of cluster 2. Citizens of cities in cluster 3 have higher quality of life than the others e.g. they feel more safe, are more committed to fight against climate change and are generally satisfied with sport facilities, beauty of the streets, public spaces and outdoor recreation. Air pollution and noise are major problems of cities in cluster 4; in this cluster, cities are not considered clean or healthy to leave in.

Table 4. Features' saliencies and clusters probabilities regarding the agree and satisfied categories

		Cluster 1	Cluster 2	Cluster 3	Cluster 4
	α	0.44	0.12	0.13	0.31
Features	Saliency				
Satisfied with sport facilities	1	0.78	0.67	0.84	0.52
Satisfied with beauty of streets	1	0.68	0.63	0.80	0.44
City committed to fight against climate change	1	0.58	0.53	0.69	0.35
Satisfied with public spaces	1	0.81	0.73	0.87	0.53
Noise is a big problem here	1	0.72	0.64	0.44	0.85
Feel safe in this city	1	0.80	0.82	0.94	0.66
Feel safe in this neighborhood	1	0.86	0.89	0.97	0.79
Administrative services help efficiently	1	0.67	0.57	0.68	0.42
Satisfied with green space	0.91	0.81	0.71	0.88	0.43
Resources are spent in a responsible way	0.88	0.53	0.49	0.63	0.35
Most people can be trusted	0.86	0.55	0.57	0.81	0.34
Satisfied with health care	0.81	0.82	0.73	0.89	0.49
Poverty is a problem	0.77	0.55	0.63	0.56	0.69
Air pollution is a big problem here	0.76	0.80	0.66	0.49	0.87
It is easy to find a good job here	0.53	0.49	0.26	0.47	0.39
This is a clean city	0.52	0.49	0.67	0.79	0.22
Satisfied with outdoor recreation	0.49	0.79	0.70	0.87	0.47
Easy to find good housing at reasonable price	0.49	0.24	0.48	0.59	0.33
City is healthy to live in	0.41	0.57	0.74	0.91	0.28
Satisfied with cultural facilities	0.32	0.95	0.91	0.57	0.68
Satisfied with public transport	0.31	0.81	0.70	0.88	0.40
Foreigner here are well integrated	0			0.39	
Presence of foreigners is good for the city	0			0.79	

6 CONCLUSIONS AND FUTURE RESEARCH

In this work, we implement an integrated EM algorithm to simultaneously select relevant features and cluster categorical data. The algorithm estimates the importance of each feature using a saliency measure. In order to test the performance of the proposed algorithm, two kinds of data sets were used: synthetic and real data sets. Synthetic data sets were used to test the ability to select the (previously known) relevant features and discard the irrelevant ones. The results clearly illustrate the ability of the proposed algorithm to recover the ground truth on data concerning the features' saliency and clustering. On the other hand, the usefulness of the algorithm is illustrated on real data based on European official statistics.

Results obtained with the data sets considered are encouraging. In the near future, an attempt to integrate both the selection of the number of clusters and of the relevant categorical features based on a similar approach will be implemented. Recently, this integration was successfully accomplished on synthetic data [16], but still offers some challenges when real data is considered.

References

[1] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research **3** (2003) 1157–1182

[2] Dy, J., Brodley, C.: Feature selection for unsupervised learning. The Journal of Machine Learning Research **5** (2004) 845–889

[3] Steinley, D., Brusco, M.: Selection of variables in cluster analysis an empirical comparison of eight procedures. Psychometrika **73**(1) (2008) 125–144

[4] Law, M., Figueiredo, M., Jain, A.: Simultaneous feature selection and clustering using mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004) 1154–1166

[5] Wallace, C., Boulton, D.: An information measure for classification. The Computer Journal **11** (1968) 195–209

[6] Dempster, A., Laird, N., Rubin, D.: Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of Royal Statistical Society **39** (1997) 1–38 Series B.

[7] Silvestre, C., Cardoso, M., Figueiredo, M.: Selecting categorical features in model-based clustering using a minimum message length criterion. In: The Tenth International Symposium on Intelligent Data Analysis - IDA 2011. (2011)

[8] Vinh, L.T., Lee, S., Park, Y.T., d'Auriol, B.J.: A novel feature selection method based on normalized mutual information. Applied Intelligence **37**(1) (2012) 100–120

[9] Constantinopoulos, C., Titsias, M.K., Likas, A. In: Bayesian Feature and Model Selection for Gaussian Mixture Models. Volume 28. IEEE Transactions on Pattern Analysis and Machine Intelligence (2006) 1013–1018

[10] Zeng, H., Cheung, Y.: A new feature selection method for gaussian mixture clustering. Pattern Recognition **42** (2009) 243–250

[11] Talavera, L.: An evaluation of filter and wrapper methods for feature selection in categorical clustering. Advances in Intelligent Data Analysis VI (2005) 440–451

[12] Dash, M., Liu, H., , Yao, J.: Dimensionality reduction for unsupervised data. In: Ninth IEEE International Conference on Tools with AI - ICTAI97. (1997)

[13] Talavera, L.: Dependency-based feature selection for symbolic clustering. Intelligent Data Analysis **4** (2000) 19–28

[14] Cover, T., Thomas, J.: 2. In: Entropy, Relative Entropy and Mutual Information. Elements of Information Theory (1991)

[15] Silvestre, C., Cardoso, M., Figueiredo, M.: Clustering with finite mixture models and categorical variables. In: International Conference on Computational Statistics - COMPSTAT2008. (2008)

[16] Silvestre, C., Cardoso, M., Figueiredo, M.: Simultaneously selecting categorical features and the number of clusters in model-based clustering. In: International Classification Conference - ICC 2011. (2011)