SELECTING CATEGORICAL FEATURES IN MODEL-BASED CLUSTERING

Cláudia M. V. Silvestre
Escola Superior de Comunicação Social, Lisboa, Portugal

Margarida M. G. Cardoso
ISCTE, Business School, Lisbon University Institute, Lisboa, Portugal

Mário A. T. Figueiredo
Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal

Keywords: Cluster analysis, Finite mixture models, Feature selection, EM algorithm, Categorical variables.

Abstract: There has been relatively little research on feature/variable selection in unsupervised clustering. In fact, feature selection for clustering is a challenging task due to the absence of class labels for guiding the search for relevant features. The methods proposed for addressing this problem are mostly focused on numerical data. In this work, we propose an approach to selecting categorical features in clustering. We assume that the data comes from a finite mixture of multinomial distributions and implement a new expectation-maximization (EM) algorithm that estimate the parameters of the model and selects the relevant variables. The results obtained on synthetic data clearly illustrate the capability of the proposed approach to select the relevant features.

1 INTRODUCTION

Clustering techniques are commonly used in several research and application areas, with the goal of discovering patterns (groups) in the observed data. More and more often, data sets have a large number of features (variables), some of which may be irrelevant for the clustering task and adversely affect its performance. Deciding which subset of the complete set of features is relevant is thus a fundamental task, which is the goal of feature selection (FS). Other reasons to perform FS include: dimensionality reduction, removal of noisy features, providing insight into the underlying data generation process. Whereas FS is a classic and well studied topic in supervised learning (i.e., in the presence of labelled data), the absence of labels in clustering problems makes unsupervised FS a much harder task, to which much less attention has been paid. A recent review and evaluation of several FS methods in clustering can be found in (Steinley and Brusco, 2008).

Most work on FS for clustering has focused on numerical data, namely on Gaussian-mixture-based methods (Constantinopoulos et al., 2006), (Dy and Brodley, 2004), (Law et al., 2004); work on FS for clustering categorical data is much rarer. In this work, we propose an embedded approach (as opposed to a *wrapper* or a *filter* (Dy and Brodley, 2004)) to FS in categorical data clustering. We work in the common framework for categorical data clustering in which the data is assumed to originate from a multinomial mixture. We assume that the number of mixture components is known and use an EM algorithm, together with an MML (minimum message length) criterion to estimate the mixture parameters (Figueiredo and Jain, 2002), (Law et al., 2004). The novelty of the approach is that it avoids combinatorial search: instead of selecting a subset of features, the probabilities that each feature is relevant are estimated. This work extends that of (Law et al., 2004), (which was restricted to Gaussian mixtures) to deal with categorical variables.

The paper is organized as follows. Section 2 reviews the EM algorithm, introduces the notion of feature saliency, and describes the proposed method. Section 3 reports experimental results. Section 4 concludes the paper and discusses future research.

M. V. Silvestre C., M. G. Cardoso M. and A. T. Figueiredo M. (2009).
SELECTING CATEGORICAL FEATURES IN MODEL-BASED CLUSTERING.
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 303-306
DOI: 10.5220/0002303203030306

2 SELECTING CATEGORICAL FEATURES

Law, Figueiredo and Jain (Law et al., 2004) developed a new EM variant to estimate the probability that each feature is relevant, in the context of (Gaussian) mixture-based clustering. That algorithm estimates a Gaussian mixture model, based on the MML criterion (Wallace and Boulton, 1968). Their approach seamlessly merges estimation, model, and feature selection in a single algorithm. Our work is based on that approach and implements a new version for clustering categorical data via the estimation of a mixture of multinomials.

Let $\underline{y} = \left\{\underline{y}_1, \ldots, \underline{y}_n\right\}$ be a sample of n independent and identically distributed random variables, $\underline{Y} = \{\underline{Y}_1, \ldots, \underline{Y}_n\}$ with $\underline{Y}_i = \{\underline{Y}_{i1}, \ldots, \underline{Y}_{iL}\}$ is a L-dimensional random variable. It is said that $\underline{Y}$ follows a K component finite mixture distribution if its log-likelihood 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 $\sum_{k=1}^{K} \alpha_k = 1$), $\underline{\theta} = (\underline{\theta}_1, .., \underline{\theta}_K, \alpha_1, .., \alpha_K)$ is the set of 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

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

2.1 EM Algorithm

The EM algorithm (Dempster et al., 1997) has been often used as an effective method to obtain maximum likelihood estimates based on incomplete data. Assuming that observed variable $\underline{Y}_i$ for $i = 1, \ldots, n$ (the incomplete data) is augmented by a cluster-label variable $\underline{Z}_i$ which is a set of K binary indicator latent variables, the complete log-likelihood is

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

Each iteration of the EM algorithm consists of two steps

- E-step: calculates the expectation of the complete log-likelihood, whit respect to the conditional distribution of $\underline{Z}$ given $\underline{y}$ under the current estimates of the parameter

$$\mathrm{E}\left[\log f(\underline{y}, \underline{Z}|\underline{\theta})|\underline{y}, \hat{\underline{\theta}}^{(t)}\right] = \log f\left(\underline{y}, \mathrm{E}\left[\underline{Z}|\underline{y}, \hat{\underline{\theta}}^{(t)}\right]|\underline{\theta}\right),$$

  where the equality results from the fact that the complete log-likelihood is linear with respect to the elements of $\underline{Z}$.

- M-step: finds the parameters which mazimize

$$\hat{\underline{\theta}}^{(t+1)} = \arg\max_{\underline{\theta}}\left(\log f\left(\underline{y}, \mathrm{E}\left[\underline{Z}|\underline{y}, \hat{\underline{\theta}}^{(t)}\right]|\underline{\theta}\right)\right)$$

2.2 The Saliency of Categorical Features

There are different definitions of feature irrelevancy; Law *et al* (Law et al., 2004) adopt the following one: a feature is irrelevant if its distribution is independent of the cluster labels, i.e., an irrelevant feature has a density which common to all clusters. The probability functions of relevant and irrelevant features are denoted by p(.) and q(.), respectively. For categorical features, p(.) and q(.) refer to the multinomial distribution. Let $\underline{B} = (B_1, \ldots B_L)$ be the binary indicators of the features, where $B_l = 1$ if the feature l is relevant and zero if irrelevant.

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 Law et al., 2004)

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

The feature saliencies are estimated using an EM variant based on the MML criterion which encourages the saliencies of the relevant features to go to 1 and the irrelevant features to go to zero, thus pruning the feature set.

2.3 The Proposed Method

We adopt the approach proposed by Law *et al* (Law et al., 2004) which is based on the MML criterion (Figueiredo and Jain, 2002). This criterion chooses the model providing the shortest description (in an information theory sense) of the observations (Wallace and Boulton, 1968). Under the MML criterion, for categorical features, the estimate of $\underline{\theta}$ is the one that minimizes the following description length function:

```
Initialization:  set the parameters of the mixture components randomly and the common distribution to cover all the data
                 set the feature saliency of all features to 0.5
                 store the initial log-likelihood and the initial message length (mindl is initialized with this value)
continue ← 1
while continue do
    while differences on likelihood are significant do
        M-step
        E-step
        if ρ_l = 1, q(y_l|θ_l) is pruned
        if ρ_l = 0, p(y_l|θ_kl) is pruned for all k
        Compute the log-likelihood and the current message length (ml)
    end while
    if ml < mindl
        mindl ← ml
        update all the parameters of the model
    end if
    if there are saliencies ∉ {0,1}
        prune the variable with the smallest saliency
    else
        continue ← 0
    end if
end while
```

Figure 1: The algorithm.

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

where c_l is the number of categories of feature l. Using a Dirichlet-type prior for the saliencies,

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

and from a parameter estimation point of view, $l(\underline{y},\underline{\theta})$ is equivalent to a posterior density. Since Dirichlet-type prior is conjugate with the multinomial, the EM algorithm to maximize $-l(\underline{y},\underline{\theta})$ is

E-step: Compute the following quantities

$$\begin{aligned} &P[Z_{ik}=1|\underline{Y}_i,\underline{\theta}] \\ &= \frac{\alpha_k \prod_{l=1}^{L}\left[\rho_l p\left(\underline{y}_{il}|\underline{\theta}_{kl}\right) + (1-\rho_l)\, q\left(\underline{y}_{il}|\underline{\theta}_{l}\right)\right]}{\sum_{k=1}^{K}\alpha_k \prod_{l=1}^{L}\left[\rho_l p\left(\underline{y}_{il}|\underline{\theta}_{kl}\right) + (1-\rho_l)\, q\left(\underline{y}_{il}|\underline{\theta}_{l}\right)\right]} \end{aligned}$$

$$u_{ikl} = \frac{\rho_l p\left(\underline{y}_{il}|\underline{\theta}_{kl}\right)}{\rho_l p\left(\underline{y}_{il}|\underline{\theta}_{kl}\right) + (1-\rho_l)\, q\left(\underline{y}_{il}|\underline{\theta}_{l}\right)} P[Z_{ik}=1|\underline{Y}_i,\underline{\theta}]$$

$$v_{ikl} = P[Z_{ik}=1|\underline{Y}_i,\underline{\theta}] - u_{ikl}$$

M-step: Update the parameter estimates according to

$$\hat{\alpha}_k = \frac{\sum_i P[Z_{ik}=1|\underline{Y}_i,\underline{\theta}]}{n}, \quad \hat{\theta}_{klc} = \frac{\sum_i u_{ikl} y_{ilc}}{\sum_c \sum_i u_{ikl} y_{ilc}},$$

$$\hat{\rho}_l = \frac{\left(\sum_{ik} u_{ikl} - \frac{K(c_l-1)}{2}\right)_+}{\left(\sum_{ik} u_{ikl} - \frac{K(c_l-1)}{2}\right)_+ + \left(\sum_{ik} v_{ikl} - \frac{c_l-1}{2}\right)_+}$$

where $(\cdot)_+$ is defined as $(a)_+ = \max\{a,0\}$.

If, after convergence of EM, all the saliencies are zero or one the algorithm stops. Otherwise, we check if pruning the feature which has the smallest saliency produces a better message length. This procedure is repeated until all the features have their saliencies equal to zero or one. At the end, we choose the model with the minimum value of $l(\underline{y},\underline{\theta})$. The algorithm is summarized in Fig. 1.

3 EXPERIMENTS

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 can have similar (but not exactly equal) distributions within the mixture components, we consider a second type of data where we simulate irrelevant features with "similar" distributions between the components. In both cases, the irrelevant

Table 1: Probabilities of the clustering five base variables categories.

	Synthetic data		Estimated parameters	
	Component 1 Number of samples: 400 $\alpha_1 = 0.4444$	Component 2 Number of samples: 500 $\alpha_2 = 0.5556$	Component 1 $\widehat{\alpha}_1 = 0.4444$	Component 2 $\widehat{\alpha}_2 = 0.5556$
Variable 1	0.7	0.1	0.6953	0.0988
	0.2	0.3	0.2013	0.3024
	0.1	0.6	0.1034	0.5988
Variable 2	0.2	0.7	0.2007	0.6936
	0.8	0.3	0.7994	0.3064
Variable 3	0.4	0.6	0.4029	0.5996
	0.6	0.4	0.5971	0.4004
Variable 4	0.5	0.49	0.4946	
	0.2	0.22	0.2049	
	0.3	0.29	0.3005	
Variable 5	0.3	0.31	0.3119	
	0.3	0.30	0.2999	
	0.4	0.39	0.3882	

features are also distributed according to a multinomial distribution. The numerical experiments refer to 8 simulated data sets. According to the obtained results using the proposed EM variant, the estimated probabilities corresponding to the categorical features almost exactly match the actual (simulated) probabilities. In Table 1 we present results which refer to one data set with 900 observations and 5 categorical variables (features). The first three variables are relevant and the last two are irrelevant, with "similar" distributions between components. The variables 1, 4 and 5 have 3 categories each and the variables 2 and 3 have 2 categories.

4 CONCLUSIONS AND FUTURE RESEARCH

In this work, we describe a feature selection method for clustering categorical data. Our work is based on the commonly used framework which assumes that the data comes from a multinomial mixture model (we assume that the number of components of the mixture model is known). We adopt a specific definition of feature irrelevancy, based on the work of (Law et al., 2004), which we believe is more adequate than alternative formulations (Talavera, 2005) that tends to discard features which are uncorrelated. We use a new variant of the EM algorithm, together with an MML (minimum message length) criterion, to estimate the parameters of the mixture and select the relevant variables.

The reported results clearly illustrate the ability of the proposed approach to recover the ground truth on data concerning the features' saliency. In future work, we will implement the simultaneous selection of features and the number of components, based on a similar approach and illustrate the proposed approach using real data sets.

REFERENCES

Constantinopoulos, C., Titsias, M. K., and Likas, A. (2006). Bayesian feature and model selection for gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1013–1018.

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

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

Figueiredo, M. and Jain, A. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:381–396.

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

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

Talavera, L. (2005). An evaluation of filter and wrapper methods for feature selection in categorical clustering. *Advances in Intelligent Data Analysis VI*, 3646:440–451.

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