Autonomous clustering characterization for categorical data

Nistor Grozavu Lazhar Labiod Younès Bennani
LIPN-UMR 7030, Université Paris 13,
99, av. J-B Clément, 93430 Villetaneuse, France
email: {firstname.secondname}@lipn.univ-paris13.fr

Abstract

This paper addresses the problem of cluster characterization by selecting a subset of the most relevant features for each cluster from a categorical dataset in an autonomous way. The proposed autonomous model is based on the Relational Topological Clustering (RTC) associated with a statistical test which allows to detect the most important variables in an automatic way without setting any parameters. The RTC approach is used to build a prototypes matrix which contains continuous variables, where each prototype vector represents correlated categorical data. Thereafter, the statistical ScreeTest is used to detect relevant and correlated features (or modalities) for each prototype. The proposed method requires simple computational techniques and the RTC topology technique is based on the principle of the self-organizing map (SOM) model. This method allows the dimensionality reduction, visualization and cluster characterization simultaneously. Empirical results based on both synthetic and real datasets from the UCI repository, are given and discussed.

1 Introduction

With the advent of high throughput technologies, dimensionality reduction has become increasingly important in data mining field. Its goal is to reduce the number of observations (samples) and to extract the most relevant information for each data [5]. Machine learning has been very successful in developing supervised and unsupervised learning algorithms for a wide range of technical applications where clustering and unsupervised feature selection are one of the most difficult tasks of this domain.

Clustering categorical data and categorical feature selection, i.e. data in which attribute values are not ordered, is a fundamental problem in data analysis. Moreover, most algorithms for clustering categorical data require the careful choice of parameter values, which makes these algorithms difficult to be used by a non-expert with the method.

To attempt the clustering visualization, the self-organizing maps and some extensions of this model were introduced in literature in the last decade [6], [3], [2]. However, the performance of those algorithms depend critically on the parameters definition, and the new challenge for the machine learning is the autonomy of the learning models.

In literature were proposed an extension of the SOM model to detect the relevant features by introducing a weighting technique called lwo-SOM [3]. Continuous weighting provides more information about the relevance of various features, and topological clustering and feature weighting are thus clearly linked. In contrast to SOM and lwo-SOM approaches, which deals with continuous data and has several parameters to set (the map size, the learning rate, weight vectors), the Relational Topological Clustering [4] allows to cluster categorical data without setting any parameter (the only parameter is for the visualization). The use of the RTC method will be the first step for our model across the autonomous adaptive learning.

An autonomous adaptive system exists in a given environment (or set of environments) and is such that it can adapt itself to those environments in an effective manner [10], [12]. The nature of this environment is an extremely important aspect of an autonomous system. In our case, one of the aspect of this environment is the nature of the data, which should be categorical data and which should contains no missed values. More details of the nature of this autonomous environment and its parameters are discussed in section 3.4.

In this paper, we consider the both cases: to reduce the data size and to eliminate the noisy features from this data. To reduce the number of observations we use the self-organization principle to build a prototype matrix which will represent the dataset. This task became more difficult, if the trained dataset is a qualitative set. To build the prototype matrix (the map), we use the Relational Toographic Clustering (RTC) method.

Feature selection is commonly used in machine learning, wherein a subset of the features available from the data are selected for application of a learning algorithm. The best
subset contains the features that give the highest accuracy score. This is an important stage of preprocessing and is one of two ways of avoiding the curse of dimensionality. The number of observations can be reduced through unsupervised learning and feature selection. The importance of each feature depends on the size of the learning dataset - for a small sample size, eliminating a relevant feature can reduce the error. Note also that irrelevant features can be very informative when used together.

In the following sections, we introduce the topological clustering based on lwo-SOM (local weighting observations) method (section 2) used for clustering characterization for continuous data. We present the proposed autonomous learning system in section 3 which allows topological clustering (section 3.2) and feature selection simultaneously (3.3). In section 4 and 5, we show experimental results obtained for several datasets. Some conclusions are discussed at the end of the paper, as are future perspectives for research in this area.

2 Topological Clustering and feature selection

Feature selection for clustering or unsupervised feature selection is used to identify the feature subsets that accurately describe the clusters. This improves the interpretability of the induced model, as only relevant features are involved in it, without degrading its descriptive accuracy. Additionally, the identification of relevant and irrelevant features with SOM learning provides valuable insight into the nature of the cluster-structure.

Feature selection for clustering analysis is difficult because, unlike supervised learning, there are no class labels for the dataset and no obvious criteria to guide the search [13]. In [2], the weight set \( \Pi \) and prototype set \( W \) provided by lwo-SOM is used to cluster the map and to select the relevant continuous features which characterize the resulting clusters associated with cells and group of cells.

The weighted SOM method were based on initial work describing the supervised model \( w-LVQ2 \). This approach adapts weights to filter the observation during the learning process. Using this model, the observations \( x \) were weighted using weight vectors \( \pi \) before computing the distance. The objective function is rewritten as follows:

\[
R_{lwo}(\chi, W, \Pi) = \sum_{i=1}^{E} \sum_{j=1}^{|W|} \Delta_{j, \chi(x_i)} \| \pi_j x_i - w_j \|^2 \quad (1)
\]

Minimization of \( R_{lwo}(\chi, W, \Pi) \) was performed by iterative repetition of the following three steps until stabilization. The initialization step determines the prototype set \( W \) and the set of associated weights \( \Pi \), at each training step \((t + 1)\). An observation \( x_i \) is then randomly chosen from the input dataset and the following operations are repeated:

- Minimize \( R_{lwo}(\chi, \hat{W}, \hat{\Pi}) \) with respect to \( \chi \) by fixing \( \hat{W} \) and \( \hat{\Pi} \). Each weighted observation \( \langle \pi_j x_i \rangle \) is assigned to the closest prototype \( w_j \) using the assignment function, defined as follows:

\[
\chi(x_i) = \arg \min_j (\| \pi_j x_i - w_j \|^2)
\]

- Minimize \( R_{lwo}(\hat{\chi}, \hat{W}, \hat{\Pi}) \) with respect to \( \hat{W} \) by fixing \( \chi \) and \( \Pi \). The prototype vectors are updated using the gradient stochastic expression:

\[
w_j(t + 1) = w_j(t) + \epsilon(t) K_{j, \chi(x_i)} (\pi_j x_i - w_j(t))
\]

- Minimize \( R_{lwo}(\hat{\chi}, \hat{W}, \hat{\Pi}) \) with respect to \( \Pi \) by fixing \( \hat{\chi} \) and \( \hat{W} \). The update rule for the feature weight vector \( \pi_j(t + 1) \) is:

\[
\pi_j(t + 1) = \pi_j(t) + \epsilon(t) K_{j, \chi(x_i)} (\pi_j(t)x_i - w_j(t))
\]

As in the traditional stochastic learning algorithm of Kohonen [6], the learning rate at time \( t \) is denoted by \( \epsilon(t) \). The training is usually performed in two phases. In the first phase, a high initial learning rate \( \epsilon(0) \) and a large neighborhood radius \( T_{max} \) are used. In the second phase, a low learning rate and small neighborhood radius are used from the beginning.

To select the relevant features associated to the most important weights, an established statistical method scree method were used on the computed weights. This method require a low computational time but deals only for the continuous data and requires some parameters to be defined as the learning rate, the map size, the weights.

Using the principle of the cluster characterization technique combined with lwo-SOM map, in the next section we present a new procedure to cluster and to select relevant categorical features in an autonomous way.

3 Autonomous cluster characterization through features selection for categorical data

Feature selection in clustering must provide features that describe the "best” homogenous cluster. Here, we used the prototype set \( PI \) provided by the RTC algorithm. We then used the variable selection approach to characterize the resulting clusters associated with cells and group of cells. Thus, to select the relevant features, we use the the Scree Test Acceleration Factor (algorithm 2).

To attempt the clustering characterization, we integrate the RTC model and variables selection schema (Scree Test) in one procedure which is presented in the algorithm 1.
3.1 The complexity of the clustering characterization procedure

Let \( n \) be the number of observations; \( d \) - the size of variables and \( C \) - the size of the map, the clustering characterization procedure is composed from three phases:

1. Clustering. Using the RTC algorithm, the complexity for this step is \( O(C \times N \times d) \);

2. Selection. The computational time of the Scree acceleration Test procedure for the \( k \) clusters is: \( O(d \log d \times C) \).

So, the total complexity time for the proposed clustering characterization technique is \( O(C \times N \times d + C \times d \log d) \). This linear complexity depends on the size of variables which is the case for all the variables selection algorithms, and on the size of the map, because the proposed method uses map prototypes to cluster and to select the relevant features.

3.2 Relational Topological Clustering

The Relational Topological clustering model have been used for multidimensional categorical data clustering and visualization. Let \( D \) be a dataset with a set \( I \) of \( N \) objects \((O_1, O_2, ..., O_N)\) described by the set \( V \) of \( M \) categorical attributes (or variables) \( V^1, V^2, ..., V^M \), each one having \( p_1, \ldots, p_m, \ldots, p_M \) categories respectively and let \( P = \sum_{m=1}^{M} p_m \) to denote the full number of categories of all variables. Each categorical variable can be decomposed into a collection of indicator variables. For each variable \( V^m \), let the \( p_m \) values to naturally correspond to the numbers from 1 to \( p_m \) and let \( V^m_1, V^m_2, ..., V^m_{p_m} \) be the binary variables such that for each \( j, 1 \leq j \leq p_m, V^m_j = 1 \) if and only if the \( V^m \) takes the \( j \)-th value. Then the data set can be expressed as a collection of \( M \) matrices \( K^m (N \times p_m) \) for \( m = 1, \ldots, M \) of general term \( k^m_{ij} \) such as:

\[
k^m_{ij} = \begin{cases} 
1 & \text{if the object } i \text{ takes the category } j \text{ of } V^m \\
0 & \text{otherwise}
\end{cases}
\]

which gives the \( N \) by \( P \) binary disjunctive matrix \( K = \{K^1, K^2, ..., K^M\} \).

To build the map, we use the RTC technique based on Relational Analysis (J. F. Marcotorchino, 2006), which allows to learn categorical datasets. The objective function to maximize in this case can be expressed using the profiles \( K_i \) of each object and the prototype \( P_l \) of each cell of the map \( C \) as following:

\[
R_{RTC}(\varphi, P) = \sum_{i=1}^{N} \sum_{l=1}^{L} K_{i(l)}^{T} (\delta(\varphi(i), l)) (\langle K_i, P_l \rangle - \alpha S_{il})
\]

\[
= \sum_{i=1}^{N} cont^T(K_i, P_{\varphi(i)})
\]

where \( S_{il} = \sum_{K_i < K_{il}} + \sum_{K_{il} > K_i} \) and

\[
cont^T(K_i, P_{\varphi(i)}) = \langle K_i, P_l^T \rangle - \alpha \sum_{l=1}^{L} K_{i(l)}^{T} \delta(\varphi(i), l) S_{il}
\]

is the regularized contribution of the object \( i \) to his winner neuron \( \varphi(i) \). We observe that the regularized contribution of the object \( i \) to \( \varphi(i) \) is a weighted sum of the contributions of \( i \) to all prototypes \( P_l (l = 1, ..., L) \) in the influence neighborhood of \( \varphi(i) \).

\[
P_{\varphi(i)}^T = \sum_{l=1}^{L} K_{i(l)}^{T} \delta(\varphi(i), l) P_l = \sum_{l=1}^{L} K_{i(l)}^{T} \delta(\varphi(i), l) \sum_{l' \in C_l} K_{l'j}
\]

is the regularized prototype of the winner neuron \( \varphi(i) \), that could be seen as a weighted sum of the prototypes \( P_l (l = 1, ..., L) \) in the influence neighborhood of \( \varphi(i) \).

In order to avoid setting of initialization parameters, we initialize the map using the classical Relational Analysis (RA) approach.

We consider here the batch SOM: the assignment step maximize the objective function by considering all prototypes \( P \) fixed, representation step maximize the same function considering the clusters set fixed (the assignment function \( \varphi \) fixed). For a fixed temperature \( T \), the maximization
occurs in two alternating phases during successive iterations. We summarize this algorithm in the following points:

Step 1. Initialization: Initialize the map $C$ using the relational analysis approach

Step 2. Assignment: The $R_{RTC}^T(\varphi, P)$ is expressed as a sum of independent terms (regularized contributions) and we can replace the both optimization problems by a set of simple equivalent problems. Indeed, $R_{RTC}^T(\varphi, P)$ can be decomposed in terms of individual contributions of each neuron, it comes:

\[
\forall i; \; \varphi(i) = \arg\max_l \text{cont}^T(K_i, P_l) \quad (7)
\]

Step 3. Maximization: The maximization phase consist in maximizing the objective function over $P$ by setting the assignment $\varphi$ in it’s constant definition. By rewriting the objective function in terms of regularized contribution of each neuron, it comes:

\[
\text{cont}^T(\varphi, P_l) = \sum_{i=1}^{N} K_i^{T}(\delta(\varphi(i), t)) \text{cont}(K_i, P_l) \quad (8)
\]

In others words, maximization step consists in updating each regularized prototype $P_l^T(t)$ of neuron $C_i$ at each iteration $t$ according to:

\[
\forall l; \; P_l^T(t) = \sum_{r=1}^{L} K_r^{T}(\delta(r, l))(t) \sum_{i' \in C_r(t)} K_{i'} \quad (9)
\]

The proposed Batch RTC algorithm is therefore as follows:

Batch RTC algorithm with a fixed T:

**Inputs**
- $C^0$: initial map with $L_{max}$ neurons. $N_{iter}$= number of iterations. $N$= number of observations. $\alpha$= similarity threshold. $K^T$= neighborhood matrix

**Initialization:** Initialize the map $C$ using RA heuristic
- Run the RA heuristic on the $K$ matrix
- Randomly place the resulting clusters on the map $C^0$
- Compute the initial prototypes

\[
\forall l; \; P_l^T(0) \leftarrow \sum_{r=1}^{L_{max}} K_r^{T}(\delta(r, l))(0) \sum_{i' \in C_r(0)} K_{i'}
\]

for $t=1$ to $N_{iter}$ do

for $i = 1$ to $N$ do {Assignment}

assign the observation $i$ to its closest neuron within the sens of contribution:

\[
\varphi(\ i) (t) = \arg \max_{l=1, \ldots, L_{max}} \text{cont}(K_i, P_l(t - 1))
\]

end for

for $l = 1$ to $L_{max}$ do {Maximization }

update prototypes according to

\[
P_l^T(t) = \sum_{r=1}^{L_{max}} K_r^{T}(\delta(r, l))(t) \sum_{i' \in C_r(t)} K_{i'}
\]

endfor

**Outputs**
- a map of $L_{max}$ cells.

3.3 Autonomous Variables Selection: Cattell Scree Test

We propose to use an established statistical method, scree test, to select the most important features [1]. This statistical test was initially developed to provide a visual technique to select eigenvalues for principal components analysis [1]. The basic idea is to generate a curve associated with eigenvalues, allowing random behavior to be identified. The number of components retained is equal to the number of values preceding this ‘scree’. Often the ‘scree’ appears where the slope of the graph changes radically. We therefore needed to identify the point of maximum deceleration in the curve.

The basic idea of scree test is to generate, for a principal components analysis (PCA), a curve associated with eigenvalues, allowing random behavior to be identified (a simple line plot). Cattell suggests to find the place where
Thus we have to process the following steps presented in
the procedure 2.

\textbf{Algorithm 2: The Scree Test Acceleration Factor}

- Input: prototype vector \( Pl \) size \( d \)
- for \( i = 1 \) to \( d \) do
  - Sort the vector in descending order \( P_{l}^{[j]} \).
  - Thus we obtain a new order
  \[
  P_{l}^{[j]} = (P_{l}^{[j],1}, P_{l}^{[j],2}, ..., P_{l}^{[j],i}, ..., P_{l}^{[j],d}) ;
  \]
  - where \( i \) indicates the index order.
- end for
- for \( j = 1 \) to \( d \) (on the sorted vector) do
  - Compute the first difference \( df_{i} = P_{l}^{[j],i} - P_{l}^{[j],i+1} \)
  - and we obtain the vector \( P_{l}^{[j]}_{df1} \)
- end for
- for \( p = 1 \) to \( d \) (on the \( P_{l}^{[j]}_{df1} \) vector) do
  - Compute the second difference \( (\text{acceleration}) \)
  \[
  acc_{i} = df_{i} - df_{i+1} \]
  - obtaining the vector \( P_{l}^{[j]}_{df2} \)
- end for
- for \( l = 1 \) to \( d \) (on the \( P_{l}^{[j]}_{df2} \) vector) do
  - Find the scree: \( \max_{i} (\text{abs}(acc_{i}) + \text{abs}(acc_{i+1})) \)
- end for

\textbf{OUTPUT:}
- Retain all the features displayed before the scree (we used the initial index values of features before sorting).

\subsection{3.4 Autonomous environment analysis}

The nature of the environment of the proposed clustering characterization method is that it deals with the categorical variables and it is not adapted for the continuous data. The used RTC and RA clustering algorithms doesn’t need the number of cluster to be fix but these methods require a parameter \( (\alpha) \) which allows to adjust the number of cluster, but usually this parameter is fixed to 0.5.

Another parameter to be fixed is the size of the map used for the visualization, but his parameter can be computed based on the data size.

The strong part of the proposed system concern the cluster characterization procedure which uses the autonomous features selection (Scree Test).

\section{4 Experimentations and validation}

\subsection{4.1 Zoo dataset:}

We use the zoo dataset to show the good performance of the proposed clustering characterization schema using the RTC algorithm. This dataset contains 101 animals described with 16 qualitative variables: 15 of the variables

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{snowman.png}
\caption{An example of the automatic scree test using a prototype vector. The axes \( X \) and \( Y \) correspond to features and prototype’s values, respectively. The scree is indicated by the vertical bar.}
\end{figure}
are binary and one is numeric with 6 possible values. Each animal is labelled 1 to 7 according to its class. Using disjunctive coding for the qualitative variable with 6 possible values, the data set consists of a $101 \times 36$ binary data matrix. All 101 animals are used for learning with a map with size $5 \times 5$ cells. The learning algorithm provides a profile prototype for each cell. At the end of the learning phase, each observation, corresponding to an animal, is assigned to the cell with the highest contribution by taking into account the neighborhood relation.

![Figure 2. A dataset with qualitative variables](image)

The RTC algorithm start with the initialization of the grid by distributing the observations using relational analysis approach. An example of the initial dataset is given in the figure 2, and it is very difficult to detect relevant features when the data contains only binary variables (0 and 1, white and black colors). But, using our proposed Clustering Characterization which allows the dimensionality reduction of the dataset, we are able to construct a prototype matrix which represents the neurons from the RTC map. This matrix contains only continuous features as it is shown in the figure 3 where the red (darkest) color corresponds to the most relevant features for the respective neuron and the blue (white) color - to the noisy features.

Using Scree Test technique for the RTC map, we will select relevant features for each cell; and we give an example of four clusters from this map: cell 1, 7, 22 and 24. The neuron 1 captured the following samples (animals): bear, boar, cheetah, leopard, lion, lynx, mole, mongoose, polecat, pussycat, raccoon. The housefly, moth and wasp characterize the 7th cell, and the neuron 22 contains: clam, crab, crayfish, lobster, starfish. Finally, the 24 micro-cluster captured these animals: frog, newt, pitviper and tuatara.

The selected features for these four cells are given in the Table 1, where 0 shows the absence of the corresponding variable (the '0' modality), and 1 - the presence of the variable. These selected features are the most relevant for each neuron which characterize each cell. These results can be easily validated by analyzing the table 1 from a zoological/biological point of view.

![Figure 3. A prototype matrix: zoo map](image)

### Table 1. Selected features on the zoo map

<table>
<thead>
<tr>
<th>Zoo</th>
<th>selected features</th>
<th>means</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell 1</td>
<td>2(1),11(1),6(0), 5(1),13(0), 9(1)</td>
<td>hair, breathes, airborne, milk, fins, toothed</td>
</tr>
<tr>
<td>cell 7</td>
<td>11(1), 12(1), 3(0), 6(1), 10(0)</td>
<td>breathes, venomous, feathers, airborne, backbone</td>
</tr>
<tr>
<td>cell 22</td>
<td>13(0), 14(5), 3(0), 6(0)</td>
<td>fins, legs, feathers, airborne</td>
</tr>
<tr>
<td>cell 24</td>
<td>3(0), 6(0)</td>
<td>Feathers, airborne</td>
</tr>
</tbody>
</table>

In order to validate the proposed method we compute also the accuracy map index before and after feature selection. For the map (size 5x5) obtained from zoo dataset the accuracy index is 89.15. The noisy features founded by the Scree Test algorithm are the features 29 and 30. We eliminate these variables from the dataset and we re-build the map and the new accuracy index increase to 95.65.

To analyze the impact of the map size on the results we build another map size 4x4, and the new accuracy index before feature selection is 84.16 but this time the noisy features are 4, 22, 29, 30, and the accuracy index after feature selection increase to 88.54.

By analyzing these indexes we can conclude that after feature selection the purity of the map are better and the smaller is the map, the bigger is the number of the eliminated features, but the accuracy index are smaller when the map size decrease.
5 Results for other datasets

We tested our proposed algorithm on additional datasets with different characteristics. For the proposed method, we show in Table 2 the feature selection results obtained for the SPECTF, nursery and Pos-Operative datasets. We will not discuss these results as it is difficult to evaluate the quality of the selected features for each cell when the intersect of the selected features for all the cells are the total number of the initial variables. The aim of these results is to show that for various datasets of different sizes we can cluster, and characterize the cells (clusters) in an autonomous way.

Table 2. Selected features on the zoo map

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Map size</th>
<th>Nb of selected features for each cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPECTF 80x1127</td>
<td>3x3</td>
<td>703, 1, 663, 864, 1011, 856, 760, 808, 772</td>
</tr>
<tr>
<td>Nursery 12960x29</td>
<td>4x4</td>
<td>15, 7, 3, 19, 14, 23, 2, 16, 7, 12, 15, 14, 12, 8, 14, 11</td>
</tr>
<tr>
<td>Pos-Operative 90x24</td>
<td>3x3</td>
<td>21, 15, 12, 7, 3, 8, 19, 20, 17</td>
</tr>
</tbody>
</table>

6 Conclusion

We have proposed in this paper a process for dimensionality reduction using features selection in the unsupervised learning paradigm in an autonomous way. This process uses the RTC algorithm to learn and to build a self-organizing map from a categorical dataset. Several experiments are given and our proposed approach demonstrated the efficiency for simultaneous clustering and feature selection. For future work, we will propose an extended model which will be able to escape the $\alpha$ parameter required by RA and RTC algorithms, and to validate the clustering characterization (feature selection for each cell) using the computed contributions during the learning step.

References


