Topological Organization for Categorical Data Clustering

Lazhar Labiod, Nistor Grozavu and Younès Bennani

Laboratoire d'Informatique de Paris-Nord
UMR CNRS 7030 Institut Galilée - Université Paris-Nord
99, avenue Jean-Baptiste Clément 93430 Villetaneuse, France
name@lipn.univ-paris13.fr

Abstract

We explore in this paper a novel topological organization algorithm for categorical data clustering and visualization named RTC. In general, it is more difficult to perform clustering on categorical data than on numerical data due to the absence of the ordered property in the data. The proposed approach is based on the self-organization principle of the Kohonen's model and uses the Relational Analysis formalism by optimizing a cost function defined as a modified Condorcet criterion. We propose an iterative algorithm, which deals linearly with large datasets, provides a natural clusters identification and allows a visualization of the clustering result on a two dimensional grid. RTC was validated on variant real datasets and the experimental results show the effectiveness of the proposed procedure.

1 Introduction

In the exploratory data analysis of high dimensional data, one of the main tasks is the formation of a simplified, usually visual, overview of data sets. This can be achieved through simplified description or summaries, which should provide the possibility to discover most relevant features or patterns. Clustering and projection are among the examples of useful methods to achieve this task. On one hand classical clustering algorithms produce a grouping of the data according to a chosen criterion. Projection methods, on the other hand, represent the data points in a lower dimensional space in such a way that the clusters and the metric relations of the data items are preserved as faithfully as possible. In this field, most algorithms use similarity measures based on Euclidean distance. However there are several types of data where the use of this measure is not adequate. This is the case when using categorical data since, generally, there is no known ordering between the feature values. In this work, we present a new formalism that can be applied to this type of data and simultaneously achieves the both tasks, clustering and visualization.
Topological learning is a recent direction in Machine Learning which aims to develop methods grounded on statistics to recover the topological invariants from the observed data points. Most of the existed topological learning approaches are based on graph theory or graph-based clustering methods.
The topological learning is one of the most known technique which allow clustering and visualization simultaneously. At the end of the topographic learning, the "similar" data will be collect in clusters, which correspond to the sets of similar observations. These clusters can be represented by more concise information than the brutal listing of their patterns, such as their gravity center or different statistical moments. As expected, this information is easier to manipulate than the original data points. The neural networks based techniques are the most adapted to topological learning as these approaches represent already a network (graph). This is why, we use the principle of the self-organizing maps which represent a two layer neural network: an entry layer and a topological layer (the map).
In order to visualize the partition obtained by the Relational Analysis approach (Marcotorchino, 2006),(Marcotorchino and Michaud, 1978) Marcotorchino proposed a methodology called "Relational Factorial Analysis"(Marcotorchino, 1991, 2000) which combines the relational analysis for clustering and the factorial analysis for the visualization of the partition on the factorial designs. It is a juxtaposition of the both methods, the methodology presented here combines the relational analysis approach and the SOM principle determined by a specific formalism to this methodology. The proposed model allows simultaneously, to achieve data clustering and visualization, indeed, it automatically provided a natural partition (i.e without fixing a priori the number of clusters and the size of each cluster) and a self-organization of the clusters on a two-dimensional map while preserving the a priori topological data structure (i.e two close clusters on the map consist of close observations in the input space). Various methods based on the principle of SOM model were proposed in the literature for bi-

nary data processing: probabilistic methods and others quantization techniques. Most of these methods operate on the data after a preliminary transformation step in order to find a continuous representation of the data, and then apply SOM model, like KACM (Cottrell and Letrmy, 2003)and the approach suggested by Leich and al (Leich, Weingessel and Dimitriadou, 1998). These methods destroy the binary nature of the data, in other words, they violate the structure of the data to meet the requirements of the method. In (Lebbah, Badran and Thiria, 2000)the authors propose BTM method (binary topological map) which operates directly on binary data based on the Hamming distance. In (Lebbah, Rogovschi and Bennani, 2007)a probabilistic version of the SOM model is proposed, based on the Bernoulli distribution adapted to the binary data (BeSOM). This paper is organized in the following way: in section 2 we present the Relational Analysis approach, and the section 3 presents the classical self-organizing maps model. We show in section 4 the formalism of the topological clustering problem in a relational framework and the proposed "Batch RTC" algorithm. The section 5 shows the experimental results and some perspectives related to the proposed approach.

2 Relational analysis approach

Relational Analysis was developed in 1977 by F. Marcotorchino and P. Michaud, inspired by the work of Marquis de Condorcet, which was interested in the 18th century with the result of collective vote starting from individual votes. This methodology is based on the relational representation (pairwise comparison) of data objects and the optimization under constraints of the Condorcet criterion.

2.1 Definitions and notations

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) $V^1, V^2., V^m, .., V^M$, each one having $p_1, .., p_m, .., p_M$ 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 $V_1^m, V_2^m, ..., V_{p_m}^m$ be the binary variables such that for each j, $1 \leq j \leq p_m$, $V_k^m = 1$ 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 $m = 1, .., M$) of general term k_{ij}^m such as:

$$k_{ij}^m = \begin{cases} 1 & \text{if the object } i \text{ takes the categorie } j \text{ of } V^m \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

which gives the N by P binary disjunctive matrix $K = (K^1|K^2|...|K^m|...|K^M)$.

2.2 Relational data representation

If the data is made up of N objects $(O_1, O_2, ..., O_N)$ on which M attributes (or variables) $(V^1, V^2, ..., V^M)$ have been measured then the "pairwise comparison principle" consists in transforming the data, which is usually, represented by a $N \times M$rectangular matrix into two squared $N \times N$ matrices S and $\bar{S}$. The matrix S, which is called the global relational Condorcet's matrix, of general term $s_{ii'}$ representing the global similarity measure between the two objects O_i and $O_{i'}$ over all the M attributes and matrix $\bar{S}$ of general term $\bar{s}_{ii'}$ which represent the global dissimilarity measure of these two objects. To get matrix S, each V^m attribute is transformed into a squared $N \times N$ matrix S^m of general term $s_{ii'}^m$ which represent the similarity measure between the two objects O_i and $O_{i'}$ with regards to attribute V^m . Then, $s_{ii'}^m = 1$ if O_i and $O_{i'}$ take the same categorie of V^m and 0 otherwise. To get matrix $\bar{S}$, a dissimilarity measure $\bar{s}_{ii'}^m$ s of objects O_i and $O_{i'}$ with regards to attribute V^m is then computed as the complement to the maximum possible similarity measure between these two objects. As the similarity between two different objects is less or equal to their self-similarities: $s_{ii'}^m \leq min(s_{ii}^m, s_{i'i'}^m)$ then $\bar{s}_{ii'}^m = \frac{1}{2}(s_{ii}^m + s_{i'i'}^m) - s_{ii'}^m$. This leads to a dissimilarity measure matrix $\bar{S}^m$. The matrices S and $\bar{S}$ are then obtained by summing, respectively, all the matrices S^m and $\bar{S}^m$, that is $S = \sum_{m=1}^{M} S^m$ and $\bar{S} = \sum_{m=1}^{M} \bar{S}^m$. The global similarity between each two objects O_i and $O_{i'}$ is thus $s_{ii'} = \sum_{m=1}^{M} s_{ii'}^m$ and their global dissimilarity is $\bar{s}_{ii'} = \sum_{m=1}^{M} \bar{s}_{ii'}^m$.
The table 1 shows different coding forms for a qualitative dataset containing 5 objects measured on a qualitative variable with 3 modalities.

	V_1
o_1	1
o_2	2
o_3	1
o_4	2
o_5	3

	V_1^1	V_1^2	V_1^3
o_1	1	0	0
o_2	0	1	0
o_3	1	0	0
o_4	0	1	0
o_5	0	0	1

	o_1	o_2	o_3	o_4	o_5
o_1	1	0	1	0	0
o_2	0	1	0	1	0
o_3	1	0	1	0	0
o_4	0	1	0	1	0
o_5	0	0	0	0	1

Table 1: Linear coding - disjunctive coding - Relational coding

2.3 Condorcet's criterion maximization

To cluster a population of N objects described by M variables, the relational analysis theory maximises the Condorcet's criterion :

$$\max_X \mathcal{R}_{RA}(S, X)$$

$$\text{with } X = \{x_{ii'}\}_{i,i':1,...,N} \quad \text{an equivalence relation defined on } I \times I.$$

Where

$$\mathcal{R}_{RA}(S, X) = \sum_{i,i'=1}^{N} s_{ii'} x_{ii'} + \sum_{i,i'=1}^{N} \bar{s}_{ii'} \bar{x}_{ii'} \quad (2)$$

$$= \sum_{i,i'=1}^{N} (s_{ii'} - \bar{s}_{ii'}) x_{ii'} + \sum_{i,i'=1}^{N} \bar{s}_{ii'} \quad (3)$$

$$= 2 \sum_{i,i'=1}^{N} (s_{ii'} - \frac{1}{2}\frac{s_{ii} + s_{i'i'}}{2}) x_{ii'} + \beta \quad (4)$$

Where $\beta = \sum_{i,i'=1}^{N} \bar{s}_{ii'}$ is a constant term, and X is the reached solution wich models a partition in a relational space (an equivalence relation), and must check the following properties:

$$\begin{cases} x_{ii} = 1, \ \forall i & \text{reflexivity} \\ x_{ii'} - x_{i'i} = 0, \forall (i, i') & \text{symmetry} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1, \forall (i, i', i'') & \text{transitivity} \\ x_{ii'} \in \{0, 1\}, \forall (i, i') & \text{binarity} \end{cases}$$

Let us consider $\mathcal{C} = \{\mathcal{C}_1, ..\mathcal{C}_l, ...\mathcal{C}_L\}$ a partition of the set I into L clusters, the Condorcet criterion breaks up into terms of contributions where the contribution $cont(i, l)$ of an object i in a cluster $\mathcal{C}_l$ of the partition is written:

$$cont(i, l) = \sum_{i' \in \mathcal{C}_l} [s_{ii'} - \alpha(\frac{s_{ii} + s_{i'i'}}{2})] \quad (5)$$

Where $\alpha \in [0, 1]$ is the similarity threshold, we have

$$\mathcal{R}_{RA}(S, X) = \sum_{i=1}^{N} \sum_{l=1}^{L} cont(i, l) \quad (6)$$

That we can express in terms of the object profile K_i representing the i^{th} row of the complete disjunctive table K and P_l the prototype of cluster $\mathcal{C}_l$, is defined in the following way:

$$s_{ii'} =< K_i, K_{i'} > \quad \text{and} \quad P_l = \sum_{i' \in \mathcal{C}_l} K_{i'} \quad (7)$$

Then, we have

$$cont(K_i, P_l) =< K_i, P_l > -\alpha S_{il} \quad (8)$$

Where $S_{il} = \frac{|\mathcal{C}_l| < K_i, K_i > + \sum_{i' \in \mathcal{C}_l} < K_{i'}, K_{i'} >}{2}$. This new formula of the contribution avoids the computation of square matrices S and $\overline{S}$ (Condorcet's matrix and its complementary) which reduces considerably the computational cost related to the contributions computation.

2.4 Relational analysis heuristic

The heuristic process consists in starting from an initial cluster (a singleton cluster) and build in an incremental way, a partition of the set I by accentuating the value of Condorcet criterion $\mathcal{R}_{RA}(S, X)$ at each assignment. We give below the description of the Relational Analysis algorithm which was used by the Relational Analysis methodology (see Marcotorchino and Michaud for further details). The presented algorithm aims at maximizing the criterion given in (4) based on the contribution computation.

Algorithm1: RA heuristic

Inputs:
L_{max}= maximal number of clusters, N_{iter}= number of iterations, N= number of examples (objects), α= similarity threshold
- take the first object as the first element of the first cluster.
- $l = 1$ where l is the current number of clusters
for t=1 to N_{iter} **do**
  for $i = 1$ to N **do**
    for $j = 1$ to l **do**
      Compute the contribution of object i : $cont(i, j)$
    end for
    $l^* = arg \max_j cont(i, j)$,
where l^* is the cluster id which has the highest contribution with the object i
    $cont(i, l^*) \leftarrow$ the computed contribution
    if $cont(i, l^*) < 0$ **and** $l < L_{max}$ **then**
      create a new cluster where the object i is the first element
      $l \leftarrow l + 1$
    else
      assign object i to cluster $\mathcal{C}_{l^*}$
    endif
  endfor
endfor
Output:
at most L_{max} clusters

We have to fix a number of iterations and the similarity threshold in order to have an approximate solution in a reasonable processing time. Besides, it is also required a maximum number of clusters, but since we don't need to fix this parameter, we put by default $L_{max} = N$. Basically, this algorithm has $O(N_{iter} \times L_{max} \times N)$ computation cost. In general term, we can assume that $N_{iter} << N$, but not $L_{max} << N$. Thus, in the worst case, the algorithm has $O(L_{max} \times N)$ computation cost.

3 Self organizing map

The model called Kohonen's self organizing map (SOM) is an artificial neural network, which learns to model a data space $(Z, z_i \in R^d)$ also called set of observations (objects) by a set of prototypes $(W, w_l \in R^d)$ (the neurons)where observations and neurons are vectors of the input space.
If the network consists of L neurons, the SOM technique provides a partition into L clusters of the input space where the number of observations $N >> L$. Each neuron l is associated with a vector of weight w_l which belongs to the input space. Thus, for a set of observations the network learns the position in this space of L centers.

For example in the trivial case where $L = N$, the best possible partition is obviously a discrete partition where each observation is isolated in a cluster (the center of each cluster corresponds to the observation forming the cluster), which minimizes the distance to all data objects.
The modelling quality depends on the used metric distance in a vector space. We use the Euclidean distance to measure the distance between an observation and a prototype (two vectors). In addition, to model inputs through prototypes, a self-organizing map $\mathcal{C}$ allows to build a graph G for structuring this space and provides a visualization in one or two dimensions of the topological links between clusters. It should be remembered that the Kohonen's network is not a simple clustering algorithm, it is a model that seeks to project multidimensional observations on a discrete space (the map $\mathcal{C}$) of small dimensions (usually 1, 2 or 3). This projection has to respect the property of "conservation" of topology of data, ie two neurons l, r which are neighbors over the discrete topological map must be associated with two close prototypes w_l, w_r compared to the Euclidean distance in the observation space.
The map $\mathcal{C}$ is in the form of an undirected graph $G = (\mathcal{C}, \mathcal{A})$, where $\mathcal{C}$ refers to the L vertices (neurons) and $\mathcal{A}$ the set of edges that gives the organization of neurons on the map $\mathcal{C}$. Thus, two neurons l, r are directly connected neighbors in the map if $a(c, r) \in \mathcal{A}$. This graph induces a discrete distance δ on the map: for any pair of neurons (l, r) of the map the distance $\delta(l, r)$ is defined as being the length of the shortest path between l and r. For every neuron l, this distance determines the Neighborhood of order d of c as following: $\mathcal{V}_c(d) = \{l \in \mathcal{C},\ \delta(c, l) \leq d\}$
This notion of neighborhood can be formalized using a kernel function $\mathbf{K}$ defined from R^+ in R^+, and decreasing such that $\mathbf{K}(0) = 1$ and $\lim_{x\to\infty} \mathbf{K}(x) = 0$ (in practice we use $\mathbf{K}(x) = e^{-x^2}$). This function generates a family of functions $\mathbf{K}^T$, defined by $\mathbf{K}^T(x) = \mathbf{K}(\frac{x}{T})$. The parameter T is analogous to a temperature, when T is hight, then $\mathbf{K}^T(x)$ remains close to 1 even for large values of x; contrarily a low value produces a $\mathbf{K}^T$ function which decreases quickly to 0. The role of $\mathbf{K}^T$ is to transform the discrete distance δ induced by the structure of the graph into a regular neighborhood parameterized by T. We will use $\mathbf{K}^T_{(\delta(l,r))}$ as a measure of effective closeness between neurons l and r. During the SOM algorithm, the value of T decreases to stabilize the solution.
The quality of the partition and topology conservation is measured using the objective function $\mathcal{R}^T_{SOM}(\varphi, W)$, which must be as low as possible.

$$\mathcal{R}^T_{SOM}(\varphi, W) = \sum_{i=1}^{N}\sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))}||z_i - w_l||^2 \qquad (9)$$

Where φ represent the assignment function such that: $\varphi(i) = l$ if $i \in \mathcal{C}_l$.

4 Relational topological clustering (RTC)

Similarly to the classical model of self-organizing map (SOM), we use for the proposed RTC model an artificial neural network with an entry layer for the observations (data) and a map $\mathcal{C}$ having a topological order for the exit. The topology of the map is defined via an undirected graph. Like the SOM algorithm, the RTC model includes the vector quantization procedure. During this procedure, each neuron of the map which is the index of a prototype for required quantization will be represented by a vector of the same dimension than the observations. Contrarily to SOM approach, quantization is done by means of assignment function φ adapted to binary data, the choice of prototypes and the assignment function is done by maximizing the objective function denoted $\mathcal{R}^T_{RTC}(\varphi, P)$. Maximization must allow on one hand, to define prototypes making possible to carry out a conservation of the data topology (defined by a measurement of contribution) and to carry out, on the other hand, a partition of set I into homogeneous sub sets.
The basic idea of the RTC approach is to maximize a new objective function defined from the classical RA criterion $\mathcal{R}_{RA}$ by adding a regularization term $\mathcal{R}_{Topo}$, which introduces a topological constraint. The RTC objective function is the follows:

$$\mathcal{R}^T_{RTC}(\varphi, \mathcal{X}) = \mathcal{R}_{RA}(S, X) + \mathcal{R}_{Topo}(\varphi, \mathcal{X}) \qquad (10)$$

Where

$$\mathcal{R}_{RA}(S, X) = \sum_{i,i'=1}^{N} \Psi_{ii'} x_{ii'} \qquad (11)$$

And

$$\mathcal{R}_{Topo}(\varphi, \mathcal{X}) = \sum_{i,i'=1}^{N} \Psi_{ii'} \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} \mathcal{X}_{i'l} \bar{\mathcal{X}}_{il} \qquad (12)$$

Where $\forall i, i'\ \ \Psi_{ii'} = s_{ii'} - \alpha(\frac{s_{ii}+s_{i'i'}}{2})$, $\mathcal{X}_{il}$ is the general term of the partition matrix $\mathcal{X}$ of set I into L clusters such that $\mathcal{X}_{il} \in \{0, 1\}$, $\sum_{l=1}^{L} \mathcal{X}_{il} = 1$, $\bar{\mathcal{X}}_{il} = 1 - \mathcal{X}_{il}$. and $\forall i, i';\ \ x_{ii'} = \sum_l \mathcal{X}_{il}\mathcal{X}_{i'l}$, which is the general term of the equivalence relation X.
This function breaks up into two terms, the first one corresponds to the Condorcet criterion $\mathcal{R}_{RA}(S, X)$ whose maximization makes possible to obtain a partition of I more compact possible within the meaning of the Condorcet criterion. The second term makes possible to take into account the influence of neighborhood between a neuron and its neighbors on the map $\mathcal{C}$. It makes possible to bring closer the partitions corresponding to two different neurons on the map in order to preserve the topological order between the various partitions. Indeed,

the second term imposes to the prototype of the neuron l to represent objects belonging to nearby neurons: if the neuron l is close to the neuron $\varphi(i)$ on the map $\mathcal{C}$, a small value $[\sum_{i'=1}^{N} \Psi_{ii'} \mathbf{K}^T_{(\delta(\varphi(i),l))} \mathcal{X}_{i'l}]$ will more penalizes the maximization of the objective function.
The temperature T adjusts the relative importance granted to both terms. Indeed, for the large values of temperature, the second term is dominating and in this case the priority is given to the topology. More T is small, more the first term is taken into account and the priority is given to the determination of prototypes representing the compact partition. The RTC approach acts in this case exactly like the Condorcean method. It is thus possible to say that the relational topological map model makes possible to obtain a regularized solution of that obtained by the Condorcean method where the regularization is obtained by the respect of the a priori topological data structure.
The development of the both terms (11) and (12) leads to the following expression of the objective function:

$$\begin{aligned}
\mathcal{R}^T_{RTC}(\varphi,\mathcal{X}) &= \sum_{i,i'=1}^{N} \Psi_{ii'} \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(l,l))} \mathcal{X}_{i'l} \mathcal{X}_{il} \\
&+ \sum_{i,i'=1}^{N} \Psi_{ii'} \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} \mathcal{X}_{i'l} \bar{\mathcal{X}}_{il} \\
&= \sum_{i,i'=1}^{N} \Psi_{ii'} \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} \mathcal{X}_{i'l} (\mathcal{X}_{il} + \bar{\mathcal{X}}_{il}) \\
&= \sum_{i,i'=1}^{N} \Psi_{ii'} \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} \mathcal{X}_{i'l} \qquad (13)
\end{aligned}$$

4.1 A new writing of the objective function

The objective function above can be expressed using the profiles K_i of each object and the prototype P_l of each cell of the map $\mathcal{C}$ as following:

$$\mathcal{R}^T_{RTC}(\varphi,\mathcal{X}) = \sum_{i=1}^{N} \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} \underbrace{\sum_{i'=1}^{N} \Psi_{ii'} \mathcal{X}_{i'l}}_{cont(i,l)} \qquad (14)$$

Replacing the contribution $cont(i,l)$ by $cont(K_i,P_l)$ gives the following writing:

$$\mathcal{R}^T_{RTC}(\varphi,P) = \sum_{i=1}^{N} \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} (< K_i, P_l > - \alpha S_{il}) \qquad (15)$$

$$= \sum_{i=1}^{N} cont^T(K_i, P_{\varphi(i)}) \qquad (16)$$

Where

$$cont^T(K_i, P_{\varphi(i)}) = \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} (< K_i, P_l > - \alpha S_{il}) \qquad (17)$$

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 $\varphi(i)$.
We can rewrite this contribution in the following simplified form:

$$cont^T(K_i, P_{\varphi(i)}) = < K_i, P^T_l > - \alpha \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} S_{il} \qquad (18)$$

Where

$$P^T_{\varphi(i)} = \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} P_l = \sum_{l=1}^{L} \mathbf{K}^T_{(\delta(\varphi(i),l))} \sum_{i' \in \mathcal{C}_l} K_{i'} \qquad (19)$$

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

4.2 RTC heuristic

In this section, we will give an algorithm suitable to the RTC's formalism. We consider here the batch SOM : the assignment step maximizes the objective function by considering all prototypes P fixed; representation step maximizes the same function considering the clusters set fixed (the assignment function φ 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 $\mathcal{C}$ using the relational analysis approach
Step 2. Assignment: The $\mathcal{R}^T_{RTC}(\varphi,P)$ 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, $\mathcal{R}^T_{RTC}(\varphi,P)$ can be decomposed in terms of individual contributions of each $i \in I$ in each cell of the map $\mathcal{C}$. It is assumed at this stage that all prototypes are fixed and remains constant by maximizing the function $\mathcal{R}^T_{RTC}(\varphi,P)$ compared to φ. It is easy to see that this maximum is reached for an assignment function defined by:

$$\forall i;\ \varphi(i) = arg \max_{l}\ cont^T(K_i, P_l) \qquad (20)$$

Step 3. Maximization: The maximization step consist in maximizing the objective function over P by setting

the assignment φ in it's constant definition. In others words, maximization step consists in updating each regularized prototype $P_l^T(t)$ of neuron $\mathcal{C}_l$ at each iteration t according to the following rule:

$$\forall l; \quad P_l^T(t) = \sum_{r=1}^{L} \mathbf{K}^T{}_{(\delta(r,l))}(t) \sum_{i' \in \mathcal{C}_r(t)} K_{i'} \qquad (21)$$

The proposed Batch RTC algorithm is presented in Algorithm2:

Algorithm2: Batch RTC algorithm with a fixed T:

Inputs

$\mathcal{C}^0$= initial map with L_{max} neurons. N_{iter}= number of iterations. N= number of observations. α= similarity threshold. $\mathbf{K}^T$= neighborhood matrix

Initialization: Initialize the map $\mathcal{C}$ using RA heuristic

- *Run the RA heuristic on the K matrix*
- *Randomly place the resulting clusters on the map* $\mathcal{C}^0$
- *Compute the initial prototypes:*

$\forall l; P_l^T(0) \leftarrow \sum_{r=1}^{L_{max}} \mathbf{K}^T_{(\delta(r,l))}(0) \sum_{i' \in \mathcal{C}_r(0)} K_{i'}$

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:

    $\varphi_{(i)}(t) = arg\max_{\{l=1,.....,L_{max}\}} cont(K_i, P_l(t-1))$

  end for

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

    update prototypes according to

    $\mathrm{P}_l^T(t) = \sum_{r=1}^{L_{max}} \mathbf{K}^T_{(\delta(r,l))}(t) \sum_{i' \in \mathcal{C}_r(t)} K_{i'}$

  endfor

endfor

Outputs

a map of L_{max} *cells.*

5 Experimentations and validation

There are many ways to measure the accuracy of clustering algorithm. One of the ways of measuring the quality of a clustering solution is the cluster purity. Let there be L clusters of the dataset I and size of cluster $\mathcal{C}_l$ be $|\mathcal{C}_l|$. The purity of this cluster is given by $\text{purity}(\mathcal{C}_l) = \frac{1}{|\mathcal{C}_l|} \max_k(|\mathcal{C}_l|_{cluster=k})$ where $|\mathcal{C}_l|_{cluster=k}$ denote the number of items for the cluster k assigned to cluster l. The overall purity of a clustering solution could be expressed as a weighted sum of individual cluster purities

$$purity = \sum_{l=1}^{L} \frac{|\mathcal{C}_l|}{|I|} purity(\mathcal{C}_l) \qquad (22)$$

In general, if the values of purity are larger, the clustering solution is better.

5.1 The datasets for validation

In this section, we evaluate the performance of the RTC heuristic on several databases available at the UC Irvine Machine Learning Repository (Asuncion and Newman, 2007)

5.2 Results on zoo dataset:

We use the zoo dataset to show the good performance of the RTC algorithm. Using disjunctive coding for the qualitative variable with 6 possible values, the data set consists of a 101×21 binary data matrix. All 101 animals are used for learning with a map size 5×5 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.

The RTC algorithm starts with the initialization of the grid by distributing the observations using Relational Analysis approach. The figure 1 shows the class of animals distributed after the initialization step of the RTC algorithm. We use the animals names used in original dataset. To visualize the coherence of the map with the labelling of animals, this figure shows the class number corresponding to each cell after the application of the majority rule in each cell. We remind that during this learning step, the neighborhood information is not considered (the neighborhood function $\mathbf{K}$ is not computed). On the initialization grid (figure 1) the observations are not well distributed, there are two set of observations labelled with 7 which are separated by 2 empty cells; we can find also four sets of animals labelled as 1 which are dispersed on the map: two sets on the left top corner, one set is situated on the left bottom corner, and the last one, on the right bottom part of the map. This map demonstrates that the classical RA doesn't use a topological information during the clustering process which could allow a better distribution of the observations. After the initialization step, the RTC algorithm will continue the learning process by taking into account the neighborhood relation between all the cells. Figure 2 shows animals names collected by each cell. The map shows that the same class of animals is assigned to cells close to each other.

We can observe that the animals corresponding to the class 1 are clustered in the cells situated on the left bottom of the map (figure 2); the birds which correspond to the class 2 are in the right bottom part of the map. Also, we can analyze that fruitbat from the class 1 situated nearest to the cell containing the birds (class 2), this is explained that the fruitbat has nearest characterization with the birds even it comes from another family. On the middle of the map there is a cell containing 2

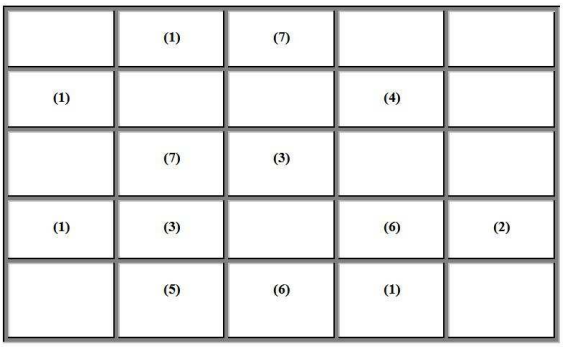

Figure 1: Initialization map using Relational Analysis algorithm

observations from two different classes: the frog (class 5) and penguin (class 2). The RTC algorithm put these two observations in the same cell because the frog and the penguin has very closest specifications even the penguin belongs to birds family and frog, from the amfibia family. Moreover, on the left of this cell there is a cell containing the animals from class 5, and on the right, a cell labelled as class 2. We have the same situation for the cell labelled as 3.5 where the toad and the tortoise has highly correlated features, and the both cells labelled as 5 and 1 are bordered on the right from this cell. The same type of analysis can be applied to the remaining clusters. To give a global view of the homogeneous clustering, we compute the clustering purity for the zoo map and we obtain a purity value of 97.84%.

We compare our map with the map obtained using the BeSOM (Bernoulli on Self-Organizing Map)(Lebbah, Rogovschi and Bennani, 2007) which use a probabilistic reformulation of the classical SOM. The map obtained using the BeSOM method is presented in the figure 3. Analyzing both maps obtained with BeSOM (figure 3) and with the proposed RTC approach (figure 2) we can detect some correlations between them: class 5 and 2 are situated in the middle of the map; the majority of the cells containing animals forming the first class are situated on the left bottom corner of the map. Comparing with the BeSOM zoo map, we can observe that RTC zoo map provides more finer cells: in the case of the BeSOM map there are three cells which contains only one observation which respectively will attribute to these ones a 100% of purity, and 8 cells containing only two observations. The RTC map has no cell which contains only one observation an has only 3 cells with two observations that means that our map has cells with a better distribution of observations. In order to show the good performance of the Relational Topological Clustering (RTC) approach we use several binary datasets of different sizes. For each dataset we learned a map of different size (from 4x4 to 10x10) and we indicate in the table 2 the purity of clustering after the first iteration using the classical RA and the

Clam Crab Crayfish lobster octopus seawasp slug starfish worm (7)	Hamster (1) Ladybird (6) Scorpion (7) (1.5)	chicken dove flamingo parakeet sparrow vulture (2)	dolphin leopard pony (1)	bass catfish chub dogfish haddock herring pike piranha stingray tuna (4)
flea honeybee moth wasp (6)	Toad (5) Tortoise (3) (3.5)	antelope hare vole (1)	gnat housefly termite (6)	mink oryx pussycat seal (1)
aardvark gorilla lion reindeer (1)	giraffe puma wallaby (1)	frog newt pitviper (5)	Frog (5) Penguin (2) (2.5)	crow hawk ostrich pheasant (2)
seasnake squirrel vampire (1)	porpoise sealion (1)	slowworm tuatara deer (3)	carp seahorse sole (4)	duck gull kiwi skua (2)
girl opossum platypus raccoon wolf (1)	boar calf elephant goat lynx (1)	bear cavy cheetah (1)	buffalo fruitbat mole mongoose polecat (1)	lark rhea skimmer swan wren (2)

Figure 2: Relational Topological Clustering : zoo database

antelope buffalo deer elephant giraffe hare mole opossum oryx vole (1)	dolphin porpoise (1)	bass catfish chub dogfish herring pike piranha stingray tuna (4)	carp haddock seahorse sole (4)	clam seawasp (7)
aardvark bear boar cheetah leopard lion lynx mink mongoose polecat puma pussycat raccoon wolf (1)	frog frog newt toad tuatara (5)	pitviper slowworm (3)	slug worm (7)	crab crayfish lobster octopus starfish (7)
calf cavy goat hamster pony reindeer (1)	scorpion (7)	kiwi ostrich penguin rhea vulture (2)	girl seal sealion (1)	platypus seasnake tortoise (3)
fruitbat squirrel vampire (1)	duck flamingo swan (2)	chicken dove lark parakeet pheasant sparrow wren (2)	flea termite (6)	gnat (6)
gorilla wallaby (1)	crow gull hawk skimmer skua (2)	honeybee wasp (6)	housefly moth (6)	Ladybird (6)

Figure 3: BeSOM map

map purity at the end of the map learning with the RTC technique. The results illustrate that the proposed technique increase the purity index compared to the classical RA and allows to obtain a topological map by computing the neighborhood function between the cells.

Table 2: Experimentation results on different datasets using RTC approach.

	DB size	Map size	RA purity	RTC purity
Zoo	101x17	5x5	69.08 %	97.84 %
Car	1728x6	10x10	70.31 %	80.17 %
Nursery	12960x8	6x6	50.47 %	78.69 %
SPECTF	267x22	4x4	57.14 %	81.82 %
Pos-Operative	90x8	5x5	71.59 %	78.21%

6 Conclusion

We have proposed in this paper a new Relational Topological model for multidimensional categorical data clustering and visualization, inspired from the SOM principle and the Relational Analysis formalism. It combines the advantages of both methods, indeed it allows a natural clusters identification without a priori fixing the number of clusters, and simultaneously provides a clusters visualization on a low dimensional lattice while preserving as faithfully as possible the topological data structure. However, this model addresses in the same way the variables describing the data, it ignores their internal structures, the number of modalities per variable and frequency of each modality. For this problem, we expect to propose a weighted model to take into account these information of variables.

Reference

Asuncion, A. and Newman, D.J. (2007). "UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine", CA: University of California, School of Information and Computer Science.

Barbara Hammer, Alexander Hasenfu, Fabrice Rossi and Marc Strickert (2007)."Topographic Processing of Relational Data". In Proceedings of the 6th Workshop on Self-Organizing Maps (WSOM 07), Bielefeld, Germany. September 2007.

Cottrell, M. and Letrmy, P. (2003). "Analyzing surveys using the Kohonen algorithm", Proc. ESANN 2003, Bruges, 2003, M.Verleysen Ed., Editions D Facto, Bruxelles, 85-92.

Forgy, E. W. (1965) Cluster analysis of multivariate data : efficiency versus interpretability of classification., in *Biometrics*, vol 21, 1965, pp768-780.

Kohonen, T. (1995) Self-Organizing Maps. *Springer Series in Information Sciences,* vol 30, Springer.

Labiod, L. (2008) *Contribution au formalisme relationnel des lassifications simultanées de deux ensembles.* (PHD thesis of Paris 6 University, 2008.)

Lebbah, M, Badran, F and Thiria, S. (2000) Topological map for binary data., in *ESANN 2000.*

Lebbah, M, Rogovschi, N and Bennani, Y. (2007) BeSOM: Bernoulli on Self-Organizing Map, in *International Joint Conference on Neural networks*, IJCNN, August 2007.

Lebbah, M, Rogovschi, N and Bennani, Y. (2008), A Probabilistic Self-Organizing Map for Binary Data Topographic Clustering, in *International Journal of Computational Intteligence and Applications* , World Scientific Publishing Compagny.p 363-383, Vol7, No.4. 2008.

Leich, F, Weingessel, A and Dimitriadou, E Competitive Learning for Binary Data., in *Proc of ICANN'98*, septembre 2-4. Springer Verlag, 1998.

Letremy, P, Traitement de données qualitatives par des algorithmes fondés sur l'algorithme de Kohonen, *SAMOS-MATISSE UMR 8595, Universit de Paris 1, 2005.*

Marcotorchino, J.F.(2006) Relational analysis theory as a general approach to data analysis and data fusion, in *Cognitive Systems with interactive sensors, 2006.*

Marcotorchino, J.F. et Michaud, P. (1978) *Optimisation en analyse ordinale des données.* (In Masson, 1978.)

Marcotorchino, J.F. (1991) *L'analyse factorielle relationnelle: partie I et II.* (Etude du CEMAP, IBM France, vol MAP-03, décembre 1991)

Marcotorchino, J.F. (2000) *Dualité Burt-Condorcet : relation entre analyse factorielle des correspondances et analyse relationnelle.* (Etude du CEMAP, IBM France, in l'analyse des correspondances et les techniques connexes. Springer 2000.)

Zighed, D. A, Hacid, H. and Aupetit, M. (2009) "Topological Learning", Proceedings of Toplearn workshop of ISMIS,Prague , 2009.