Dimensionality reduction for binary data

Nistor Grozavu, Lazhar Labiod, Younès Bennani

LIPN CNRS UMR 7030, Paris 13 University
99, J.-B. Clément, 93430 Villetaneuse, France
Name.Surname@lipn.univ-paris13.fr

Abstract

This paper addresses the problem of selecting a subset of the most relevant features for each cluster from a binary dataset. The proposed model is based on the Relational Topological Clustering (RTC) associated with a statistical test which allows to detect the most important variables. 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 for each prototype. This method allows the dimensionality reduction, visualization and cluster characterization simultaneously. The first results using this technique 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. 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 main objectives of dimensionality reduction are thus (Grozavu et al. 2009):

- to facilitate the visualization and data understanding;
- to reduce the required storage space;
- to reduce the learning time;
- to identify the relevant features.

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. Several methods can be used to reduce the size of features.

- Selection: a subset of features is chosen from the initial data space;
- Transformation: new features are built in a transformed space - an output space.

In this work we are interested in feature selection based techniques wich allow to detect the relevant features during the learning process for each neuron (prototype).

2 Feature selection and cluster characterization

In this section we propose to use an automatic procedure to select the relevant features using the prototype matrix obtained during the learning process. After obtaining the map we apply the variables selection procedure to detect the relevant features for each cell separately.

Relational Topological Clustering

For building the map, we use the RTC technique based on Relational Analysis (J. F. Marcotorchino, 2006), which allow to learn qualitative datasets.

The RTC objective function in terms of regularized contribution of each neurone is:

$$cont^T(\varphi, Pl) = \sum_{i=1}^{N} \mathbf{K}^T_{(\delta(\varphi(i),l))} cont(K_i, Pl) \qquad (1)$$

The prototypes (neurons) are computed using the following expression:

$$\forall l; \;\; Pl^T(t) = \sum_{r=1}^{L} \mathbf{K}^T(_{\delta(r,l))}(t) \sum_{i' \in \mathcal{C}_r(t)} K_{i'} \qquad (2)$$

where $Pl^T(t)$ is the computed prototype of neuron $\mathcal{C}_l$ at each iteration t.

2.1 Automatic Variables Selection : Catell Scree Test

We propose to use an established statistical method, *scree test*, to select the most important features (Cattell, 1966).

This statistical test was initially developed to provide a visual technique to select eigenvalues for principal components analysis (Cattell, 1966). 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.

Figure 1 shows an example of a curve generated using a prototype vector. We observed the scree on the 19th feature which means that the irrelevant features have index values lying in the range $[20-40]$. We used an automated process to apply this technique to each weight vector $\mathbf{P}l_j = (Pl_j^1, Pl_j^2, ..., Pl_j^d)$.

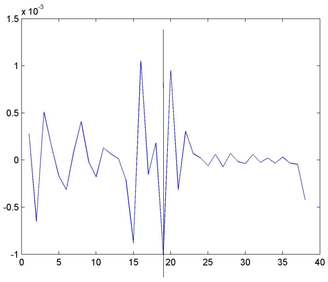

Figure 1: An example of the automatic scree test using a prototype vector. The axes X and Y correspond to features and prototype's values, respectively. The scree is indicated by the vertical bar.

Thus we have to process the following steps presented in the procedure 1.

2.1.1 Complexity of the Scree Test procedure

The Scree Test acceleration procedure has four steps until finding the scree in the vector. We will analyze all these steps:

- To made the sort of the weight vector we are using the Merge sort procedure which has an algorithmic complexity: $O(d \log d)$;

Algorithme 1: The Scree Test Acceleration Factor

```
Input: prototype vector $Pl$ size $d$
for $i = 1$ to $d$ do
    Sort the vector in descending order $Pl^{[j]}$.
    Thus we obtain a new order
    $Pl^{[j]} = (Pl_{\cdot}^{[j],1}, Pl_{\cdot}^{[j],2}, ..., Pl_{\cdot}^{[j],i}, ..., Pl_{\cdot}^{[j],d})$ ; where $i$
    indicates the index order.
end for
for $j = 1$ to $d$ (on the sorted vector) do
    Compute the first difference $df_i = Pl_{\cdot}^{[j],i} - Pl_{\cdot}^{[j],i+1}$ and
    we obtain the vector $Pl_{df1}^{[j]}$
end for
for $p = 1$ to $d$ (on the $Pl_{df1}^{[j]}$ vector) do
    Compute the second difference (acceleration)
    $acc_i = df_i - df_{i+1}$ obtaining the vector $Pl_{df2}^{[j]}$
end for
for $l = 1$ to $d$ (on the $Pl_{df2}^{[j]}$ vector) do
    Find the scree: $\max_i \, (abs(acc_i) + abs(acc_{i+1}))$
end for
OUTPUT:
Retain all the features displayed before the scree (we
used the initial index values of features before sorting).
```

- The complexity for the first difference df_i is the $O(d)$;
- For the second difference the complexity is the same as previously: $O(d)$;
- Even if the scree is in the beginning of the vector, the algorithm must look over entire weight vector to see if there is no another bigger scree, and the complexity is $O(d)$.

As there is no nested loops, the total computational time for the Scree Test acceleration algorithm is the sum of the complexity of the four steps, and for a weight vector it is $O(d \log d + 3d)$.

2.2 Automatic cluster characterization trough features 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 (Wiratunga et al. 2006). Feature selection in clustering must provide features that describe the "best" homogenous cluster. Here, we used the prototype set Pl and prototype provided by the RTC algorithm. We then used the 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 1).

Algorithme 2: The Clustering Characterization Procedure

```
Input: Dataset $X$ size $N \times d$
for $i = 1$ to $n$ do
    Build a topological map size $C$ using the RTC algorithm
end for
for $j = 1$ to $|C|$ (for each prototype) do
    Find the relevant subset of features using the ScreeTest
    procedure (for each cell of the cluster)
end for
OUTPUT: The relevant subset of variables characterizing
the $C$ clusters of the map.
```

To attempts the clustering characterization, we integrate the RTC model and variables selection schema (Scree Test) in one procedure which is presented in the algorithm 2.

2.2.1 The complexity of the clustering characterization procedure

Let n be the number of observations; m -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/logarithmic 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 are used the map prototypes to to cluster and to select the relevant features.

3 Experimentations and validation

3.1 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 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 101×36 binary data matrix. All 101 animals are used for learning with a map with 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.

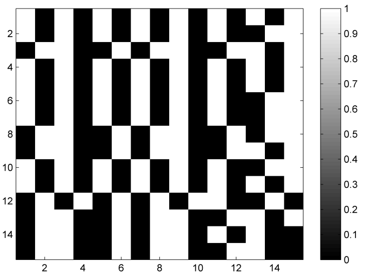

Figure 2: A dataset with qualitative variables

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.

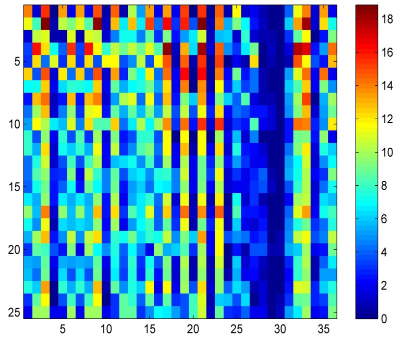

Figure 3: A prototype matrix: zoo map

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 this data: 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.

Table 1: Selected features on the zoo map

Zoo	selected features	means
cell 1	2(1), 11(1), 6(0), 5(1), 13(0), 9(1)	hair , breathes, airborne, milk, fins, toothed
cell 7	11(1), 12(1), 3(0), 6(1), 10(0)	breathes , venomous, feathers, airborne, backbone
cell 22	13 (0), 14 (5), 3 (0), 6 (0)	fins, legs, feathers, airborne
cell 24	3 (0), 6 (0)	Feathers, airborne

4 Conclusion

We have proposed in this paper a process for dimensionality reduction using features selection in the unsupervised learning paradigm. This process uses the RTC algorithm to learn and to build a self-organizing map from a qualitative dataset. We described a first testing using a statistical method for autonomous unsupervised feature selection. Our approach demonstrated the efficiency for simultaneous clustering and feature selection.

References

R.Cattell, 1996. The scree test for the number of factors. Multivariate Behavioral Research ,1:245276, 1966.

Grozavu N., Bennani Y. and M. Lebbah (2009). *From variable weighting to cluster characterization in topographic unsupervised learning.* IJCNN'09: Proceedings of the 2009 international joint conference on Neural Networks, isbn 978-1-4244-3549-4, pages 609–614, Atlanta, Georgia, USA.

Guérif S. and Y. Bennani (2007). *Dimensionality reduction trough unsupervised features selection.* International Conference on Engineering Applications of Neural Networks, Hellas.

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

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

Strickert M., Sreenivasulu N., Peterek S., Weschke W., Mock H.-P. and U. Seiffert (2006). *Unsupervised Feature Selection for Biomarker Identification in Chromatography and Gene Expression Data.* In ANNPR, pages 274-285.

Wiratunga N., Lothian R. and S. Massie (2006). *Unsupervised Feature Selection for Text Data.* In ECCBR, Lecture Notes in Computer Science, v. 4106, pages 340-354.