On Finding Complementary Clusterings

Timo Pröscholdt and Michel Crucianu

CEDRIC - Conservatoire National des Arts et Métiers
292 rue St Martin, 75141 Paris Cedex 03 - France

Abstract. In many cases, a dataset can be clustered following several criteria that complement each other: group membership following one criterion provides little or no information regarding group membership following the other criterion. When these criteria are not known *a priori*, they have to be determined from the data. We put forward a new method for jointly finding the complementary criteria and the clustering corresponding to each criterion.

1 Introduction

Consider, for example, a large set of images of blue and silver Mercedes and Toyota cars. Here, color and brand are two categorical variables that complement each other in describing the car images. Suppose that neither the variables nor their values are known *a priori*, but each image is represented by several automatically extracted low level visual features. Can one discover, by analyzing this data, the presence of two complementary categorical variables, each of them having two possible values? This would allow, for example, to improve image database summarization and to automatically find relevant search criteria.

We address this problem for data in a vector space, by looking for *complementary* clusterings in subspaces of the full space. Two clusterings of a same dataset are complementary if cluster membership according to one clustering provides little or no information regarding cluster membership according to the other. Each clustering corresponds to a categorical variable, with each cluster representing one different "value" of that variable. We assume here that for each categorical variable there is a linear subspace of the full space where the data points group in such a way that each cluster is one value of that variable. To *separate* these variables, we further consider that they should be independent on the available data. Obviously, not every dataset will show such *combinatorial* structure, where each cluster in the full space is the intersection of clusters found in different subspaces. Also, automatically found clusterings may not correspond to "meaningful" categorical variables (like color or brand).

To find arbitrarily oriented subspaces with complementary clusterings (see e.g. Fig. 1) we consider derived variables and group them into disjoint subsets on the basis of their mutual information (MI). Since we aim to find complementary *clusterings*, we compute the entropy (used for measuring MI) on a clustering of the projected data and add cluster quality to the optimization criterion.

The next section briefly reviews some existing work that can be related to the problem we aim to solve. Our method for finding complementary clusterings is described in Section 3. The evaluation in Section 4 on a synthetic dataset and on a real database shows that this method can produce good results.

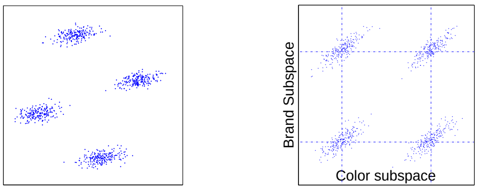

Fig. 1: Original data (left) and separated complementary clusterings (right).

2 Related work

Subspace clustering has received significant interest, motivated by cases where data points are related in different ways in different subspaces of the description space (see [1]). These methods aim to jointly uncover clusters and subspaces of the data space where each cluster is best described. A few recent methods attempt to find complementarity relations between the resulting subspaces. The proposal in [2] relies on an iterative algorithm performing at each iteration a clustering followed by the selection of the most discriminant subspace using linear discriminant analysis (with the labels resulting from the clustering) and then by a projection on the orthogonal residual subspace. An appropriate choice of the number of clusters appears to be critical. Two different methods are suggested in [3]. One extends the k-means algorithm in order to simultaneously perform clustering in several subspaces and make these subspaces orthogonal. The other method improves the fitting of a factorial density model with an extension of Expectation-Maximization. Both methods require prior knowledge of the number of subspaces and of the number of clusters in each subspace.

To identify complementary categorical variables, we could look for statistically independent components in the data. Several methods for performing independent subspace analysis were proposed, including multidimensional independent component analysis (MICA) [4], tree-dependent component analysis (TCA) [5] and independent variable group analysis (IVGA) [6] (see [7] for an agglomerative version, AIVGA). TCA aims to find a linear transform and a tree-structured graphical model such that the transformed data fits the model well. Each connected subgraph in the model (each tree in the forest) defines an independent subspace. TCA is quite general but has a high computational complexity. IVGA attempts to determine subsets of initial variables describing the data such that mutually dependent variables are grouped together while mutually independent (or weakly dependent) ones end up in separate subsets. IVGA and AIVGA have a lower complexity than TCA, but only consider the initial variables. The methods for finding independent subspaces cannot directly address our goal of identifying complementary clusterings, because they consider data points and not clusters. This often leads to subspaces where data has highly skewed rather than multi-modal distributions, which does not support clustering. Furthermore, two subspaces that are independent at the level of clusters can be dependent at the level of data points (see e.g. Fig. 1, right).

3 Proposed method

To be able to find complementary clusterings in arbitrarily oriented subspaces, the method suggested below is built upon TCA [5]. Consider $x = (x_1, \ldots, x_m)^\top$ a multivariate random variable in $\mathbb{R}^m$ having the probability distribution $p(x)$. TCA is looking for a model of $p(x)$ that consists in an invertible matrix $\mathbf{W}$ and a tree $T(\mathcal{V}, \mathcal{E})$ of vertices $\mathcal{V}$ and edges $\mathcal{E}$ such that, for $s = \mathbf{W}x$, the distribution of s factorizes in T, i.e. $p(s) = \prod_{i \in \mathcal{F}} p(s_i) \prod_{j \in \mathcal{U}} p(s_j | s_{\pi_j})$, $\mathcal{F}$ being the *founder* nodes and $\mathcal{U} = \mathcal{V} \backslash \mathcal{F}$. If T is a forest (i.e. has several disconnected trees) rather than a spanning tree, then each tree of the forest corresponds to an independent subspace. To allow for such cases, a penalty for dense forests is employed. Let $D(p||q)$ denote the Kullback-Leibler divergence between two pdfs p and q, $I(x_u, x_v) = D(p(x_u, x_v)||p(x_u)p(x_v))$ be the pairwise MI between two variables x_u and x_v and $I(x_1, \ldots, x_m) = D(p(x)||p(x_1) \cdots p(x_m))$ be the m-fold MI between the components of x. It is shown in [5] that the best TCA model for $p(x)$ is found by minimizing

$$J(x, \mathbf{W}, T) = I(s_1, \ldots, s_m) - \sum_{(u,v) \in \mathcal{E}} I(s_u, s_v) \tag{1}$$

with respect to $\mathbf{W}$ and T. But $p(x)$ is unknown, so the criterion in (1) must be replaced by an empirical contrast function. Since $I(s_1, \ldots, s_m) = \sum_u H(s_u) - H(s)$, $H(s) = H(x) + \log|\det \mathbf{W}|$, $I(s_u, s_v) = H(s_u) + H(s_v) - H(s_u, s_v)$ and $H(x)$ does not depend on $\mathbf{W}$ or T, an appropriate contrast function is

$$J_{MI} = \sum_u \hat{H}(s_u) - \log|\det \mathbf{W}| - \sum_{(u,v) \in \mathcal{E}} \left[\hat{H}(s_u) + \hat{H}(s_v) - \hat{H}(s_u, s_v)\right] \tag{2}$$

For TCA, the entropies in (2) are obtained by estimating (1D and 2D) data density using e.g. kernels. To optimize J_{MI} with respect to $\mathbf{W}$ and T, the algorithm proposed in [5] alternates minimization steps with respect to each variable, the other being fixed. To minimize with respect to T, a greedy algorithm for the maximum weight forest problem is employed (see [5]). If $s = \mathbf{W}x$ is whitened, then the minimization with respect to $\mathbf{W}$ can be performed by iteratively selecting pairs of indices (i, j) and allowing the rows i, j to vary (rotations in the spanned subspace) while keeping fixed all the other rows. To avoid poor local minima, a coarse exhaustive search is first performed, then refined optimization by gradient descend starts from the coarse minimum.

The proposed method follows the TCA approach but, to find complementary clusterings rather than generic independent subspaces, it has two important changes (see also Algorithm 1). First, cluster membership for a clustering in one subspace should provide as little information as possible regarding cluster membership for a clustering in another subspace. So, while the MI is still given by (2), the entropies are estimated on clustered data; the resulting cost is denoted J_{MIC}. Second, only a good clustering can be interpreted as a categorical variable (with each cluster corresponding to one value of that variable). So, a clustering quality term J_{CQ} is added to the MI term J_{MIC} in the contrast function

$$J_{CC} = J_{MIC} + \alpha J_{CQ} \tag{3}$$

with α controlling its impact. Together with the clustering algorithm, the J_{CQ} measure employed for clustering quality should avoid overfitting and penalize poor clustering. The choice of a value for α will depend on the clustering algorithm and on the corresponding selection of J_{CQ}. Note that the value of α is not critical if the dataset does show good complementary clusterings.

Algorithm 1 Identification of complementary clusterings.

```
Require: dataset $\mathbf{X}$, dense forest penalty $w_0$, step size $\phi$, threshold $\theta$
  initialize newScore $= \infty$, $\mathbf{W}$: random with whitened $s = \mathbf{W}x$
  repeat
    Score = newScore
    for each pair $(i, j)$, $j < i$, of rows of $\mathbf{W}$ do
      for each rotation of angle $k\phi$, $k \in \{1, \ldots, \frac{180}{\phi}\}$, of rows $(i, j)$ of $\mathbf{W}$ do
        find $T$ using greedy maximum weight forest algorithm
        compute and store $J_{CC}(k)$
      end for
      $J^*_{ij} = \min_k J_{CC}(k)$
      if $J^*_{ij} <$ newScore then
        newScore $= J^*_{ij}$
        $\mathbf{W} = \arg J^*_{ij}$
      end if
    end for
  until Score $-$ newScore $< \theta$
  perform final clustering in each resulting subspace
  return transformed data $\mathbf{S} = \mathbf{WX}$, newScore, complementary clusterings
```

To estimate the entropies in J_{MIC}, clustering has to be performed at every step of the optimization in algorithm 1, so a fast solution is needed. A simple grid-based method was employed here; it considers adjacent bins having a density above a threshold ρ to be part of the same cluster. Bin size and ρ allow to control complexity, so J_{CQ} can simply count the number of clusters in 1D and 2D.

The use of clustering raises several issues with respect to the optimization process. First, during the coarse exhaustive search by rotations in the subspace spanned by rows i, j, the number of clusters obtained for the 1D projections is likely to vary. To make the corresponding values of MI (computed on clustered data) comparable, they have to be normalized; for every coarse rotation angle, the value computed for MI is divided by the maximal value that could be obtained with the same numbers of clusters. Also, the contrast function is now discontinuous when $\mathbf{W}$ varies, so gradient descent can no longer be applied; refined search is like coarse search, but with a smaller step size. A careful analysis shows that the complexity of Algorithm 1 is $O(m^3 n)$, where m is the original dimension of the data and n the number of data points.

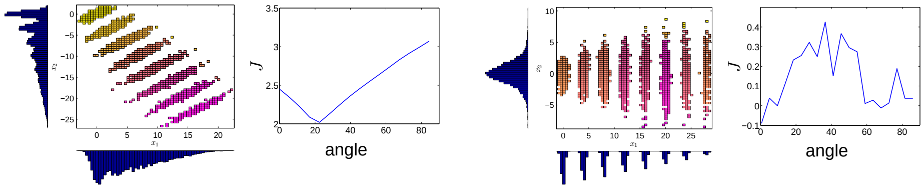

Fig. 2: Result and cost function for TCA (left) and proposed method (right).

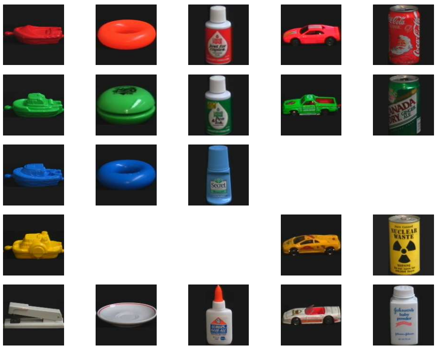

Fig. 3: Images of the 21 objects in the reduced COIL database. Objects on a same line have similar color, objects on a same column have similar shape.

4 Experimental evaluation

This method was tested on computer-generated data and also on a real dataset. Fig. 2 shows a comparison, on a simple 2D example, between TCA [5] and our method that aims to find complementary clusterings. TCA outputs two very skewed components, while the method introduced here finds one component showing good clustering and another component on which the projected data is Gaussian. Even in this simple case the contrast function is not smooth and has several local minima, so the coarse exhaustive search cannot be avoided.

The real dataset we considered is a subset of 21 classes (called below "reduced COIL" database) selected from COIL-100[1] so as to reveal two rather complementary categorical variables, one corresponding to object color and the other to shape (see Fig. 3). Each class of COIL-100 has 100 images of the same object taken from different angles. The global description of each image, combining color, texture and shape information, is represented by a 7-dimensional vector. The basic example given in the introduction is inspired by this dataset. As shown in Fig. 4, the proposed method is able to find two complementary clusterings that correspond to the two categorical variables, i.e. object shape (with 5 possible values) and respectively color (also 5 possible values).

[1] `http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php`

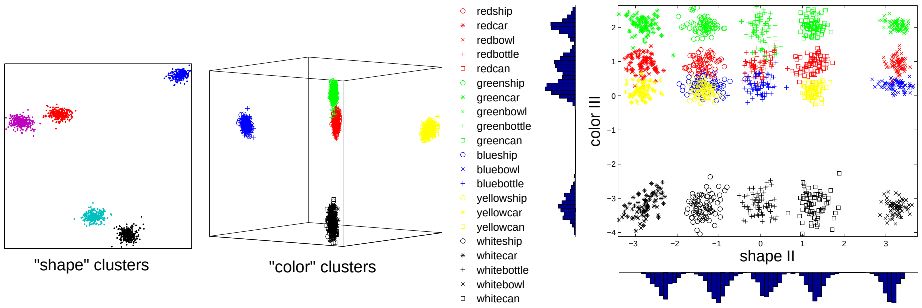

Fig. 4: Results on the reduced COIL database: the 2 complementary clusterings (left and middle, high MI between 1D projections within each subspace) and combinatorial structure along 2 complementary dimensions (right, very low MI).

5 Conclusion

Finding complementary categorical variables that describe a set of data (if such variables exist) can be useful for e.g. summarizing image databases or revealing relevant search criteria. To address this problem, we suggest a method that aims to find arbitrarily oriented subspaces showing complementary clusterings, without knowing the expected number of subspaces or of clusters within each subspace. This method is based on tree-dependent component analysis but, to find complementary clusterings rather than generic independent subspaces, the mutual information is estimated on clustered data and a clustering quality term is included in the contrast function. The experiments show that the proposed method does output the desired results on simple synthetic data and on a real dataset. Future work should address the scalability to high-dimensional data.

References

[1] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

[2] Ying Cui, Xiaoli Z. Fern, and Jennifer G. Dy. Non-redundant multi-view clustering via orthogonalization. In *ICDM'07: Proc. 7th IEEE Intl. Conf. on Data Mining*, pages 133–142, Washington, DC, USA, 2007. IEEE Computer Society.

[3] Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. Simultaneous unsupervised learning of disparate clusterings. *Stat. Anal. Data Min.*, 1(3):195–210, 2008.

[4] Jean-François Cardoso. Multidimensional independent component analysis. In *Proc. ICASSP'98*, pages 1941–1944, Seattle, WA, USA, 1998.

[5] Francis R. Bach and Michael I. Jordan. Beyond independent components: trees and clusters. *J. Mach. Learn. Res.*, 4:1205–1233, 2003.

[6] Esa Alhoniemi, Antti Honkela, Krista Lagus, Santeri Jeremias Seppä, Paul Wagner, and Harri Valpola. Compact modeling of data using independent variable group analysis. *IEEE Trans. Neural Networks*, 18(6):1762–1776, 2007.

[7] Antti Honkela, Jeremias Seppä, and Esa Alhoniemi. Agglomerative independent variable group analysis. *Neurocomputing*, 71(7-9):1311–1320, 2008.