Clayton copula and mixture decomposition

Etienne Cuvelier and Monique Noirhomme-Fraiture

FUNDP(Facultés Universitaires Notre-Dame de La Paix)
Institut d'Informatique
B-5000 Namur, Belgium
(e-mail: cuvelier.etienne@info.fundp.ac.be,
noirhomme.monique@info.fundp.ac.be)

Abstract. A symbolic variable is often described by a histogram. More generally, it can be provided in the form of a continuous distribution. In this case, the problem is to solve the most frequent problem in data mining, namely: to classify the objects starting from the description of the variables in the form of continuous distributions. A solution is to sample each distribution in a number N of points, and to evaluate the joint distribution of these values using the copulas, and also to adapt the dynamical clustering (nuées dynamiques) method to these joint densities. In this paper we compare the Clayton copula and the Normal copula for more than 2 dimensions, and we compare results of clustering by using on the one hand the method based on the Clayton copula and traditional methods (MCLUST, and K-means). Our comparison is based on 2 well-known classical data files.
Keywords: symbolic data analysis, mixture decomposition, Clayton copula, clustering.

1 Introduction

The mixture decompostion is a classical tool used in clustering. The method consists in estimating a probability density function from a given sample in R^q, considering that the reached function f is a finite mixture of K densities:

$$f(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot f(x_1, ..., x_q, \beta_i) \tag{1}$$

with $\forall\ i \in \{1, ..., K\}$, $0 < p_i < 1$, and $\sum_{i=1}^{K} p_i = 1$. The function $f(., \beta)$ is a density function with parameter β belonging to R^d and p_i is the probability that one element of the sample get the density $f(., \beta_i)$. In this clustering approach each component of the mixture corresponds to a cluster.
To find the partition $P = (P_1, ..., P_K)$, which is the best adapted to the data two main algorithms were proposed : the EM algorithm (Estimation, Maximisation) [Dempster *et al.*, 1977] and the dynamical clustering algorithm [Diday *et al.*, 1974].
A use of the dynamical clustering algorithm in the symbolic data analysis framework when the data are distribution probabilities was proposed in [Diday, 2002]. In a symbolic data table, a statistical unit can be described by

numbers, intervals, histograms and probability distributions. We suppose to have a table T with n lines and p columns, and that the j^{th} column contains probability distributions, i.e. if we note Y^j the j^{th} variable then Y_i^j is a distribution $F_i(.)$ for all $i \in \{1, ..., n\}$. To cluster this last type of data two main ideas were proposed in [Diday, 2002]. The first idea is to use as sample the values of the distributions found in table T in q quite selected values $T_1, ..., T_q : \{(F_i(T_1), ..., F_i(T_q)) : i \in \{1, ..., n\}\}$. The second idea is to estimate the margins of $f(., \beta_i)$ in a first step , and to join them in a second step using copulas.
[Vrac *et al.*, 2001] used this approach with success to cluster atmospheric data with the Franck copula of dimension 2 (i.e. with only two real values T_1 and T_2 where distributions are computed). The starting point of our work is to extend this approach with copulas with a higher number of dimensions with the Clayton n-copula.
The organization of the paper is as follows. In section 2 we set the symbolic data analysis framework for the mixture decomposition when data are probability distributions. A general presentation of the copulas is made in section 3, and we focus in section 4 on the Clayton copula. In the following section we show the implementation and results, and we conclude with perspectives and future work in the last section.

2 The symbolic data analysis framework

2.1 Distributions of distributions

We suppose to have a table T with n lines and p columns, and that the j^{th} column contains probability distributions, i.e. if we note Y^j the j^{th} variable then Y_i^j is a distribution $F_i(.)$ for all $i \in \{1, ..., n\}$. In the following we note ω_i the concept described by the i^{th} row, and $F_{\omega_i}(.)$ the associated distribution. We choose q real values $T_1, ..., T_q$ (we don't discuss of the choice of this values here), and for each $i \in \{1, ..., n\}$ we compute $F_{\omega_i}(T_1), ..., F_{\omega_i}(T_q)$. Then, if we call Ω the set of all concepts, the joint distribution of the $F_i(T_j)$ values is defined by:

$$H_{T_1,...,T_q}(x_1, ..., x_q) = P\left(\omega \in \Omega : \{F_\omega(T_1) \leq x_1\} \cap ... \cap \{F_\omega(T_q) \leq x_q\}\right) \quad (2)$$

which is called distribution of distributions. The classical classification method consists in considering this distribution as the result of a finite mixture distributions:

$$H_{T_1,...,T_q}(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i) \quad (3)$$

with $\forall\ i \in \{1, ..., K\} : 0 < p_i < 1$ and $\sum_{i=1}^{K} p_i = 1$.
The distribution of i$^{\text{th}}$ cluster is given by $H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i)$, where parameter $\beta_i \in R^d$, and p_i is the probability that one element is in this cluster.

If we take a look at the densities, then the probability density of H is

$$h(x_1, ..., x_q) = \frac{\partial^q}{\partial x_1 ... \partial x_q} H(x_1, ..., x_q) \tag{4}$$

And the mixture densities is given by:

$$h_{T_1,...,T_q}(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot h^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i) \tag{5}$$

2.2 Clustering algorithm

The clustering algorithm proposed by [Diday, 2002] is an extension of the dynamical clustering method [Diday *et al.*, 1974] for density mixtures. The main idea is to estimate at each step, the density which describes at best the clusters of the current partition P, according to a given quality criterion. We considered the classifier log-likelihood :

$$lvc(P, \beta) = \sum_{i}^{K} \sum_{\omega \in P_i} log(h(w)) \tag{6}$$

where

$$h(w) = h_{T_1,...,T_q}(F_\omega(T_1), ..., F_\omega(T_q)) \tag{7}$$

The classification starts with a random partition, then the two following steps are repeated:

- **Step 1 : Parameters estimation**
  Find the vector $(\beta_1, ..., \beta_K)$ which maximizes the chosen criterion;
- **Step 2 : Distribution of units in new classes**
  Build new classes $(P_i)_{i=1,...,K}$ with parameters found at Step 1 :

$$P_i = \{\omega : p_i \cdot h(\omega, \beta_i) \geq p_m \cdot h(\omega, \beta_m) \forall m\} \tag{8}$$

until the stabilization of partition.

2.3 Estimation

Before using this algorithm we must know how to estimate the density of each cluster.
For univariate distributions we may use :

- a parametric approach, and use well-known laws as the Beta law (Dirichlet's law in one dimension) or the Normal law,

- a non-parametric approach, as the kernel density estimation :

$$\hat{f}(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{9}$$

  where
  - $(X_1, ..., X_n)$, is the sample over which the estimation is made,
  - K is the kernel density function (many possible choices...)
  - h is the window width, and can be automatically estimated with Mean Integrated Square Error(MISE) formulae $h = 1.06\sigma N^{-1/5}$ [Silverman, 1986], where σ is the standart deviation of the sample.

For multivariate distributions, we can also use parametric estimation, with a Normal multivariate distribution for example like in [Fraley and Raftery, 2002] or a non parametric approach (the kernel estimation exists also in higher dimensions, but is heavier in calculations), but we can also attempt to re-build the joint distributions H with marginals coupling, by using copula, and at the same time have a model of the dependence structure of the data.

3 Multivariate copulas

A multivariate copula, also called n-copula, is a function C from $[0,1]^n$ to $[0,1]$ with the following properties :

- $\forall\ \mathbf{u} \in [0,1]^n$,
  - $C(\mathbf{u}) = 0$, if at least one coordinate of $\mathbf{u}$ is 0,
  - $C(\mathbf{u}) = u_k$, if all coordinates of $\mathbf{u}$ are 1 except u_k

- $\forall\ \mathbf{a}, \mathbf{b} \in [0,1]^n$, such that $\mathbf{a_i} \leq \mathbf{b_i}, \forall\ 1 \leq i \leq n$,

$$V_C([\mathbf{a}, \mathbf{b}]) \geq 0, \tag{10}$$

  where $[\mathbf{a}, \mathbf{b}] = [\mathbf{a_1}, \mathbf{b_1}] \times ... \times [\mathbf{a_n}, \mathbf{b_n}]$, and $V_C([\mathbf{a}, \mathbf{b}])$ is the nth order difference of H on $[\mathbf{a}, \mathbf{b}]$:

$$V_C([\mathbf{a}, \mathbf{b}]) = \Delta_{\mathbf{a}}^{\mathbf{b}} C(\mathbf{t}) = \Delta_{a_n}^{b_n} \Delta_{a_{n-1}}^{b_{n-1}} ... \Delta_{a_2}^{b_2} \Delta_{a_1}^{b_1} C(\mathbf{t}) \tag{11}$$

  with

$$\Delta_{a_k}^{b_k} C(\mathbf{t}) = C(..., t_{k-1}, b_k, t_{k+1}, ...) - C(..., t_{k-1}, a_k, t_{k+1}, ...) \tag{12}$$

The copulas are powerfull tools in modeling dependences since Abe Sklar stated the following theorem [Sklar, 1959]:
Let H be an n-dimensional distribution function with margins $F_1, ..., F_n$. Then there exists an n-copula C such that for all $\mathbf{x}$ in $\bar{R}^n$,

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)). \tag{13}$$

If $F_1, ..., F_n$ are all continuous, then C is unique; otherwise, C is uniquely determined on Range of $F_1 \times ... \times$ Range of F_n.

In fact the copula captures the dependence structure of the distribution. In our case, if we note a univariate margin :

$$G_T(x) = Pr\left(\omega \in \Omega : \{F_\omega(T) \leq x\}\right) \tag{14}$$

then the mixture can be written as follows

$$H_{T_1,...,T_q}(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot C^i(G^i_{T_1}(x_1), ..., G^i_{T_q}(x_q); \beta_i) \tag{15}$$

and in terms of densities

$$h^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i) = \prod_{i=1}^{q} \frac{dG^i_{T_i}}{dx}(x_i) \times \frac{\partial^q}{\partial u_1 ... \partial u_q} C^i(G^i_{T_1}(x_1), ..., G^i_{T_q}(x_q); \beta_i) \tag{16}$$

The use of copulas allows us to estimate all the marginals first, and in a second time to estimate the parameters of each copula. The copula modelises the dependences of the $F_\omega(T_i)$ values inside each cluster. Note well that this use of copulas can be made, not only when the original data are symbolic data described by a continuous distribution, but also with quantitative unspecified variables.

4 Clayton copula

In the following we present the Clayton's copula we use for our implementation, and the Normal copula for comparison.

The Clayton copula is an Archimedean copula. These copulas are generated by a function ϕ, called the generator:

$$C(u_1, ..., u_n) = \phi^{-1}\left(\sum_{i=1}^{n} \phi(u_i)\right) \tag{17}$$

where ϕ is a function from $[0, 1]$ to $[0, \infty]$ such that:

- ϕ is a continuous strictly decreasing function
- $\phi(0) = \infty$
- $\phi(1) = 0$
- ϕ^{-1} is completely monotonic on $[0, \infty[$ i.e.

$$(-1)^k \frac{d^k}{dt^k} \phi^{-1}(t) \geq 0 \tag{18}$$

for all t in $[0, \infty[$ and for all k.

If we use $\phi_\theta(t) = t^\theta - 1$ as generator, then we get the Clayton's copula

$$C(u_1, ..., u_n) = \left(1 - n + \sum_{i=1}^{n} u_i^{-\theta}\right)^{-1/\theta} \tag{19}$$

which is a copula only if $\theta > 0$.
We choose this copula in the set of the multivariate Archimedean copulas because as showed in [Cuvelier and Noirhomme-Fraiture, 2003], the density is easy to compute:

$$c(u_1, ..., u_q) = \left(1 - q + \sum_{i=1}^{q} u_i^{-\theta}\right)^{-q-\frac{1}{\theta}} \prod_{j=1}^{q} \left(u_j^{-\theta-1} \{(j-1)\theta + 1\}\right). \tag{20}$$

It is important to notice that all the k-margins of an Archimedean copula are identical: $C(u_1, ..., u_{n-1}, 1) = \phi^{-1}\left(\sum_{i=1}^{n-1} \phi(u_i)\right)$. This fact limits the nature of dependence structure in these families because it introduces a certain symmetry.
The Normal copula is built by the most obvious process: the inversion method. If we have a multivariate distribution H, with margins $F_1, ..., F_n$, then for any $\mathbf{u}$ in $[0, 1]^n$:

$$C(u_1, ..., u_n) = H(F_1^{(-1)}(u_1), ..., F_n^{(-1)}(u_n)) \tag{21}$$

is a copula. Let ρ be a positive correlation matrix, Φ_ρ the Normal multivariate distribution defined with this matrix, and Φ the standard Gaussian distribution. The Normal copula is then defined by:

$$C(u_1, ..., u_n) = \Phi_\rho(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)). \tag{22}$$

and its density is given by

$$c(u_1, ..., u_n) = \frac{1}{|\rho|^{\frac{1}{2}}} \, exp\left(-\frac{1}{2} \, \varsigma^\tau(\rho^{-1} - I) \, \varsigma\right) \tag{23}$$

where $\varsigma = \Phi^{-1}(u_i)$, and $\mathbf{I}$ is the $(n \times n)$ unity matrix. This copula has two main advantages : there is a formula to calculate its density in any dimension and, more significantly, a large set of parameters ($\frac{n \cdot (n-1)}{2}$) which indicates that one can have a very flexible modelisation of the dependence.
To show the difference between these two copulas, we generated 1000 random couples of numbers, once with Clayton copula ($\theta = 5$, figure 1), and then with the Normal copula (with a correlation of 0.5 between the two variables, figure 2).
As we can see the spatial distributions of the generated points have radically different forms. That implies that the choice of one of these two copulas

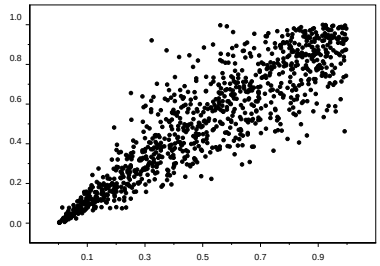

Fig. 1. Dependence structure of Clayton copula

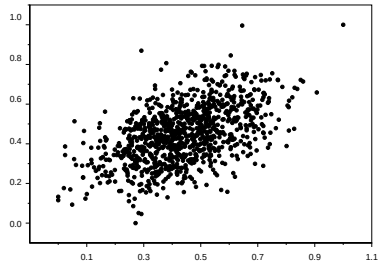

Fig. 2. Dependence structure of normal copula

will influence the shape of the clusters we can retrieve in the data. The Normal copula, and more generally the Normal distribution, tends to form elliptic groups whereas, as we can see, the copula of Clayton will tend to form groups "with pear shape".
In fact the "pear shape" shown in figure 1 is due to a property of the Clayton copula called **lower tail dependence**: a copula C has lower tail dependence if

$$lim_{u \to 0} \frac{C(u,u)}{u} > 0 \tag{24}$$

Of course the use of the Normal copula, in addition with Normal margins, corresponds to the use of the Normal multivariate distribution which was already largely studied and used in clustering methods. We will compare our results to the results of MCLUST [Fraley and Raftery, 2002] on the same data set.

5 Implementation of the algorithm and results

In this section we call our clustering algorithm (i.e. the dynamical clustering algorithm, with Clayon copula): Clayton Copule-Based clustering (CCBC). We compare the results of CCBC to the k-means implemented in S-Plus, and to the Model-Based clustering (MCLUST, [Fraley and Raftery, 2002] and [Fraley and Raftery, 1999]).
Our implementation of CCBC was made in the statistical language S, using the S-plus software. To estimate the unidimensionnal margins, we used kernel density estimation for margins (with Normal kernel).
To test our implementation we used two classical data sets. We used first the

-	**CCBC**	MCLUST	k-means
Misclass. Numb.	**9**	5	17
Percent.	**6%**	3.33%	11.33%

Table 1. Misclassified data from Fisher's Iris

very well known Iris database from Fisher. The data set contains 3 classes of 50 instances each, where each class refers to a type of Iris plant (Iris Setosa, Iris Versicolour and Iris Virginica). The 4 numerical attributes are : sepal length, sepal width, petal length and petal width. We found the same clusters with few misclassified individuals as it can be seen in table 1. The results are encouraging, especially taking into account the fact that MCLUST uses multivariate Normal laws, and so uses 6 parameters for each law, which supposes a greater flexibility to adapt it to various dependence structures.
After this we used the UCI Wisconsin diagnostic breast cancer data. In a widely publicized work [Mangasarian *et al.*, 1995], 176 consecutive future cases were successfully diagnosed from 569 instances through the use of linear programming techniques to locate planes separating classes of data. Their results were based on 3 out of 30 attributes: `extreme area`, `extreme smoothness` and `mean texture`. The three explanatory variables were chosen via cross-validation comparing methods using all subsets of 2, 3, and 4 features and 1 or 2 linear separating planes. The data is avalaible from the UCI Machine Learning Repository (http://kdd.ics.uci.edu/). The three variables of interest are shown in figure 3, and we can see that, if the joint distribution of variables `extreme smoothness` and `mean texture` seems Normal, on the other hand, the two other joint distributions are closer to Clayton copula.
We can see in table 2 that the mixture model with the Clayton copula captures the structure dependence of the breast cancer data better than the multivariate Normal distribution, in spite of the fact that all the k-margins of the copula of Clayton are identical, i.e. that this one seeks clusters necessarily presenting a certain symmetry.

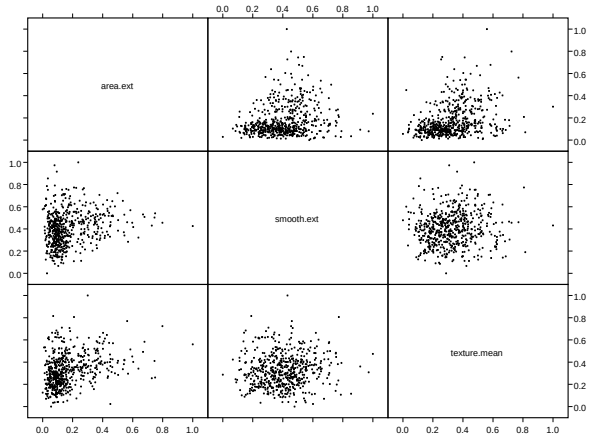

Fig. 3. Pair plots of Wisconsin Diagnostic Breast Cancer Data

-	**CCBC**	MCLUST	k-means
Misclass. Numb.	**27**	29	62
Percent.	**4.7%**	5%	10.89%

Table 2. Misclassified data from Wisconsin Diagnostic Breast Cancer Data

6 Conclusions

Mixture decomposition is a tool for classification which has already largely proved its reliability. In the same way the interest of the Copulas in the study of the dependence structures is well-known. One of the main interest of the copulas is to escape to the normality assumption and to the linear correlation.

We have shown that we can obtain equivalent or better results for clustering as with other methods, even if Clayton copula shares the weakness of all the Archimedean copulas [Nelsen, 1999]: first, in general all the k-magins of an archimedean n-copula are identical, secondly, the fact that there are only one or two parameters limits the nature of the dependence structure in these families. To overcome this weakness, in future work we intend to use other copulas with more flexible dependence structures. Now we can start to test CCBC on symbolic data.

References

[Cuvelier and Noirhomme-Fraiture, 2003]Etienne Cuvelier and Monique Noirhomme-Fraiture. Mélange de distributions de distributions, décomposition

de mélange avec la copule de clayton. In *XXXV èmes Journées de Statistiques*, 2003.

[Dempster *et al.*, 1977]A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society (Series B)*, 39(1):1–38, 1977.

[Diday *et al.*, 1974]Edwin Diday, A. Schroeder, and Y. Ok. The dynamic clusters method in pattern recognition. In *IFIP Congress*, pages 691–697, 1974.

[Diday, 2002]E. Diday. Mixture decomposition of distributions by copulas. In *Classification, Clustering and Data Analysis*, pages 297–310, 2002.

[Fraley and Raftery, 1999]C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysis. Technical report, Department of Statistics, University of Washington, 1999.

[Fraley and Raftery, 2002]C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[Mangasarian *et al.*, 1995]Olvi L. Mangasarian, W. Nick Street, and William H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43:570–577, 1995.

[Nelsen, 1999]R.B. Nelsen. *An introduction to copulas*. Springer, London, 1999.

[Silverman, 1986]B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

[Sklar, 1959]Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications Statistiques Université de Paris*, 8:229–231, 1959.

[Vrac *et al.*, 2001]Mathieu Vrac, Edwin Diday, Alain Chédin, and Philippe Naveau. Mélange de distributions de distributions, décomposition de mélange de copules et application à la climatologie. In *Actes du VIIIème congrés de la Société Francophone de Classification*, pages 348–355, 2001.