Multi-facet determination for clustering with Bayesian networks

Fernando Rodríguez-Sánchez, Pedro Larrañaga, Concha Bielza

Department of Artificial Intelligence, Universidad Politécnica de Madrid, Madrid, Spain

Abstract

Real world applications of sectors like industry, healthcare or finance usually generate data of high complexity that can be interpreted from different viewpoints. When clustering this type of data, a single set of clusters may not suffice, hence the necessity of methods that generate multiple clusterings that represent different perspectives. In this paper, we present a novel multi-partition clustering method that returns several interesting and non-redundant solutions, where each of them is a data partition with an associated facet of data. Each of these facets represents a subset of the original attributes that is selected using our information-theoretic criterion UMRMR. Our approach is based on an optimization procedure that takes advantage of the Bayesian network factorization to provide high quality solutions in a fraction of the time.

1 Introduction

Clustering is a fundamental tool for data exploration that finds interesting patterns by grouping objects based on some notion of similarity. Traditional clustering algorithms aim at finding the "best" solution at partitioning the data. However, real-world applications often involve multifaceted data where several reasonable interpretations are allowed. This is especially true for high-dimensional domains, where instances can often be grouped based on different purposes. For example, consider a collection of face images, which can be organized based on their pose or identity, or a collection of web pages, which can be classified taking into consideration their structure, content, inbound hyperlinks, etc.

When clustering this type of data, a single partition may not suffice, hence the necessity of methods that generate multiple clustering solutions. One of the first approaches that comes to mind when searching for several plausible clusterings is the naïve application of various clustering algorithms [32, 25]. However, this approach, while conceptually easy to understand, has two big disadvantages [2]. First, there is a difficulty in quantitatively evaluating the degree of similarity between the generated solutions, and second, there is an inability to know which and how many algorithms need to be applied. To avoid these issues, multi-partition clustering (MPC) methods generate multiple clustering solutions that segments the data using a collection of clusters and an array of features.

In this paper, our goal is to induce multiple facets of data that can be meaningfully clustered. To accomplish this, we propose an optimization-based approach that selects several non-redundant feature sets and generates a clustering solution for each of them. For our method to find multiple clustering solutions in different (and possibly overlapping) subsets, the relevancy of the subset's features is maximized while its redundancy with previously chosen partitions is minimized. Finally, the combination of this information-theoretic facet determination process with a model-based clustering algorithm motivate us on using Bayesian networks, whose factorization improve the interpretability and quality of the MPC solution.

1.1 Related work

MPC algorithms can be classified according to the way new partitions are found: sequentially or simultaneously.

Sequential MPC algorithms (also referred to as alternative clustering algorithms) retrieve new partitions from data that are distinct to the previously generated ones. Gondek & Hofmann [21] proposed a method that finds an alternate clustering by maximizing the pairwise mutual information between the new clustering variable and the attributes, conditioned on previous clustering. Subsequently, Bae & Bailey [2] developed the COALA algorithm, which, given a known cluster solution, applies an agglomerative clustering algorithm in combination with a series of pairwise cannot-link constraints. These constraints are imposed to objects that, in the previous solution, belong to the same cluster. [14] posteriorly presented NACI, an information-theoretic approach that maximizes the mutual information between the new clustering and data instances, while it minimizes the mutual information between the new and reference clusterings. These three methods are only able to produce a single alternative solution, however, there may exist more than two plausible groupings. Following this line of work, Cui, Fern & Dy [12, 13] developed a general purpose framework that iteratively searches for alternative clustering solutions in subspaces that are orthogonal to previously found ones. Davidson & Qi [17, 42] also presented two approaches that, independently of the specific clustering algorithm being used, are able to sequentially generate various alternative solutions. In the first one, new clustering solutions are generated by transforming previously used data with a distance function that has been learned according to a set of must-link and cannot-link constraints. The second one improves previous methodby minimizing the Kullback-Leibler divergence between the original and transformed data distributions. Both methods can be classified in the constrained optimization paradigm. Finally, Niu et al. [38] proposed a spectral clustering algorithm that generates alternative solutions by minimizing the Hilbert-Schmidt Independence Criterion.

Ont he other side, simultaneous MPC algorithms generate new partitions without taking into consideration previous ones. Caruana et al. [8] first formulated an approach that is able to generate a set of potentially interesting solutions by either randomly initializing the clustering algorithm or by using random feature weights. This collection of solutions is subsequently grouped using an agglomerative clustering based on the pairwise similarity between solutions. Jain et al. [26] then proposed two MPC algorithms: a k-means variant that returns decorrelated clusterings based on the notion of orthogonality, and a sum of parts approach, which models the clustering problem as one of learning the component distributions when data has been sampled from a convolution of mixture distributions. A model-based clustering method called CAMI was presented by [15], which simultaneously uncovers a pair of clusterings that maximizes the data likelihood and minimizes the mutual information between them. [16] induced multiple suboptimal spectral clustering solutions by using each eigenvector. Both [22] and [37] introduced a nonparametric Bayesian model that is able to discover multiple clustering solutions and the feature subsets that are relevant to each of them. In a more closely related way to this paper, Chen et al. [11], Liu et al. [34] and Poon et al. [41] advocated for the use of probabilistic graphical models in simultaneous multi-partition clustering. In the first one, an advanced greedy search algorithm called EAST hill-climbs the space of regular hierarchical latent class models (HLCMs) [48] using five search operators. A much faster approach for learning HLCMs is proposed in [34], where the authors formulate a feature clustering approach that keeps grouping variables with high pairwise mutual information until a stopping threshold is surpassed. Poon et al. [41] proposes an extension of the Gaussian mixture model, called Latent Pouch Model, and a score-based greedy search algorithm composed of seven operators to learn it.

Other type of methods that try to solve the problem of clustering multi-faceted data are those of

ensemble and multi-view clustering. Ensemble clustering creates a series of diverse base clusterings and then combines them to produce a unified solution. There are different sources of diversity in the ensembles, the main ones are non-identical feature subsets, different set of data instances and different clustering algorithms [46]. Multi-view clustering searches for complementary subspaces or feature subsets and then combines the clustering solutions for each of them into a single one. Three main techniques can be distinguished: subspace clustering [29], multiple kernel clustering [49] and co-clustering [3]. The main distinction between these two approaches and multi-partition clustering resides in their application of the *consensus principle*, which aims at maximizing the agreement between multiple distinct views, thus generating a clustering solution that is influenced by all of them [30].

1.2 Contributions of this work

Learning from these approaches to the MPC paradigm, we propose a novel method that is able to recover alternative clustering solutions based on a facet determination process. Similar to the works of [12, 13, 17, 38, 42], we address the generation of alternative clusterings in an iterative manner where relevant but distinct partitions are retrieved. In contrast to these methods, our approach is not based on data transformations or orthogonal subspaces, it extends the idea of unsupervised feature selection by determining several meaningful subsets of attributes that are alternate to each other, known as facets. Our work follows the ideas presented in [11], [34] and [41] about systematically identifying several data perspectives, clustering the data along each one and presenting the results to the domain experts for their selection. However, even though we all follow this concept while using probabilistic graphical models, our methods differ in the identification of these facets and their posterior partition construction.

Our facet determination process is lead by an information-theoretic criterion called unsupervised-maximum-relevancy-minimumm-redundancy (UMMR) that maximizes the subset's relevancy and minimizes its average redundancy with previously learned partitions. Its relevancy constraint takes inspiration on the work of Feng et al. [20], which pleads on maximizing the overall subset's information while minimizing the subset's internal redundancies. Our work follows this lead but doesn't approximate the joint entropy as the sum of its individual entropies or penalize the facet's internal redundancies with a sum of pairwise mutual information values. Our function achieves an optimal calculation of the joint entropy, that is both honest to the joint probability distribution and reasonable to compute, by factorizing this distribution with a Bayesian network. This factorization is also key in calculating Watanabe's total correlation [47], which measures the amount of information being shared between the facet's features. On the other side, to quantify the alternativeness between facets we calculate the Normalized Mutual Information [31] between them. This allows us to iteratively generate new facets that are non-redundant given they are considered to be independent between them. This calculation is also benefited by the Bayesian network's factorization, greatly reducing the number of parameters to be measured.

This process is approached as an optimization procedure where our defined criterion corresponds to a single objective function that is optimized using evolutionary algorithms. This type of algorithms suit us perfectly given the combinatorial aspect of the facet discovery problem. To take advantage of our chosen optimization scheme, we rely on constraint-based Bayesian network learning methods, which make independence tests that are then reused inside the optimization algorithm. Finally, once a facet has been determined as optimal, a model-based clustering algorithm based on Bayesian networks is applied, which takes the facet's network as an starting point and forms the new partition. This work presents a novel approach to the alternative clustering problem that is able to fully exploit the Bayesian network factorization benefits in each of its steps.

1.3 Organization of this paper

The rest of this paper is organized as follows. In section 2 we introduce some preliminaries about information theory and its relationship with statistical independence. In section 3, we define our information-theoretic criterion for facet determination. In section 4, we discuss how we approach the combinatorial problem of selecting multiple relevant and non-redundant subsets. In section 5, we present the routine of transforming an interesting facet into a partition by describing our model-based clustering algorithm. Finally, in section 6 we report our conclusions and observances.

2 Preliminaries

2.1 Information theory and statistical independence

Although initially designed to better understand communication channels, Shannon's information theory [44] has been applied to many and diverse fields in which pattern recognition is included. This theory provides measures to amount the information of the probability distributions associated with random variables.

The *entropy* of a random variable is the fundamental unit of information. It measures the average amount of information required to describe it (its uncertainty). Let X be a discrete random variable with a vector $\mathcal{X}$ of possible values, its formula can be defined as follows:

$$H(X) = -\sum_{x \in \mathcal{X}} P(x)\ logP(x) \tag{1}$$

where $H(X)$ denotes the entropy of X and $P(x)$ represents the probability density function (PDF) of X for the value x, given that $P(x) = P(X = x),\ x \in \mathcal{X}$. While the base of the logarithm can vary, Shannon's information theory uses the value of 2 as standard. This formula can be further defined for more than one discrete variable. For two discrete random variables X and Y with their joint PDF $P(x, y)$, their *joint entropy* is defined as:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y)\ logP(x, y) \tag{2}$$

When the value of X is known, the *conditional entropy* amounts the uncertainty left in Y given the information of X. The conditional entropy is less than or equal to the joint entropy of both variables, and it is only equal to the entropy when both variables are independent:

$$H(X|Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y)\ logP(x|y) \tag{3}$$

The notion of independence is of central importance in probability theory. We say that two variables X and Y are independent, denoted as $X \perp Y$, if $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$. By following this definition we can illustrate the relationship between the joint and the conditional entropy with the following formula:

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{4}$$

There is a concept in information theory that strongly relates with statistical independence, it is the *mutual information* (MI). MI is introduced by Shannon to quantify the amount of information a couple of variables X and Y share. MI is always positive and equal to zero if and only if both variables are statistically independent.

$$I(X; Y) = I(Y; X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y)\ log\frac{P(x, y)}{P(x)P(y)} \tag{5}$$

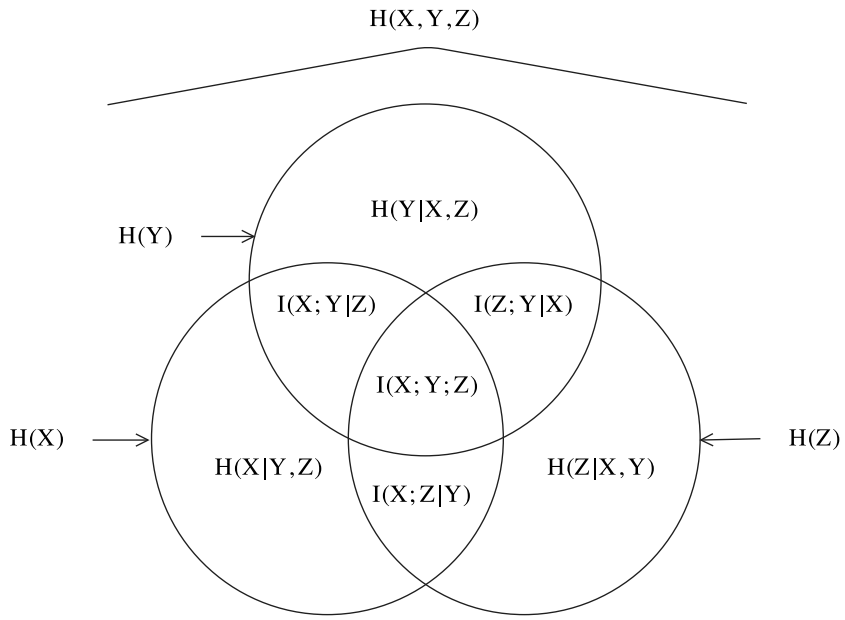

Figure 1: *A Venn diagram illustrating the relationhips between information-theoretic measures of the joint distribution of X, Y and Z. The surface area corresponds to the associated measure's quantity. This illusrtation is inspired by [27].*

When related to the concept of Entropy, MI can be seen as the amount uncertainty about a variable that can be reduced by knowing the value of another one:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) - H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned} \tag{6}$$

However, independence is not a notion, two variables could be independent until evidence about a third one is obtained. For example, let Z be a third discrete random variable with a vector of values $\mathcal{Z}$, we say that X and Y are conditionally independent given Z, denoted $X \perp Y|\ Z$, if and only if $P(X,Y|Z) = P(X|Z)\ P(Y|Z)$. There is an information-theoretic concept that relates to the notion of conditional independence, the *conditional mutual information* (CMI). The CMI measures how the uncertainty of a variable is affected by knowing information about a third one. Equivalently to MI, CMI is always positive and equal to zero if and only if both variables are conditionally independent.

$$I(X;Y|Z) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} P(x,y,z)\ log \frac{P(z)P(x,y,z)}{P(x,z)P(y,z)} \tag{7}$$

CMI also relates to the concept of Entropy with the following equalities:

$$\begin{aligned} I(X;Y|Z) &= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \\ &= H(X|Z) + H(Y|Z) - H(X,Y|Z) \\ &= H(X|Z) - H(X|Y,Z) \end{aligned} \tag{8}$$

3 Discovery of alternative partitions by facet determination

Let $D = \{d_1, ..., d_N\}$ be a series of N data objects with an associated discrete feature array of $F = \{X_1, ..., X_l\}$. Given a group of partitionings $P_{curr} = \{P_1, ..., P_t\}$ retrieved from D, our goal is to find a new partition P_{t+1} that is relevant to the domain expert and novel when compared to the ones inside P_{curr}. Each of these partitions is represented by a set of clusters $\{C_1, ..., C_c\}$ and a subset $S \in F$ of the original features. As discussed in Section 1.1, there are many ways to define alternative partitions. Our proposal is based on a facet determination process that extends the notion of unsupervised feature selection by aiming to find multiple subsets of features $\{S_1, ..., S_v\}$, such that each of them contributes a unique and interesting view of data. Once a desirable facet has been determined, the partition is fully constituted by applying a clustering algorithm that finds its best set of clusters. This procedure will iteratively search for alternative partitions until a number v of them has been reached. We define an information-theoretic criterion that maximizes the relevancy of the facets and minimizes the redundancy between them. This criterion is the basis of our partition process.

3.1 Unsupervised Maximum Relevancy Minimum Redundancy (UMRMR)

3.1.1 Maximum relevancy

In terms of information theory, the purpose of unsupervised feature selection is to find a feature set $S \in F$ with m features that maximally preserves the information of F. This is equivalent to maximizing their MI, which can be expressed with the subsequent equation where X_{s_m} $(1 \leq m \leq l)$ represents any of the original features:

$$max\ I(S; F) = max\ I(\{X_{s_1}, ..., X_{s_m}\}; \{X_1, ..., X_l\}) \tag{9}$$

From this equation we can define the subsequent lemma:

Lemma 3.1. *Let F be the full set of features and let S be the selected subset of features. If all the features belonging to S are present in F, then the following equality holds: $max\ I(S; F) = max\ H(S)$*

Proof 1. *Since the definition of mutual information is $I(S; F) = H(S) - H(S|F)$ and the definition of conditional entropy is $H(S|F) = H(S, F) - H(F)$, we can substitute the original formula for $I(S; F) = H(S) + H(F) - H(S, F)$. Now, given that all the provided information by the variables present in S is already present in F, we can substitute $H(S, F) = H(F)$, resulting: $I(S; F) = H(S) + H(F) - H(F) = H(S)$*

However, while this criterion assures that only the maximally informative features are selected, it is likely they have rich dependencies among them, given it doesn't penalize repeated information. As a result, we penalize this value using Watanabe's *total correlation* [47]. For a given set of n features $\{X_1, ..., X_n\}$, the total correlation $TC(X_1, ..., X_n)$ amounts the information being shared among all the attributes that belong to it. Total correlation is a multivariate generalization of the MI that measures the variables' grade of dependence. It is equivalent to the Kullback-Leibler divergence from the joint dstribution $P(X_1, ..., X_n)$ to the independent distribution of $P(X_1) \cdots P(X_n)$, which can be reduced to a difference of their entropies:

$$TC(X_1, ..., X_n) \equiv D_{KL}[P(X_1, ..., X_n)||P(X_1) \cdots P(X_n)] = \sum_{i=1}^{n} H(X_i) - H(X_1, ..., X_n) \tag{10}$$

This quantity is always positive, and its value is zero if and only if the variables are independent. This means that the information of one of them doesn't provide any clue about the information of the rest.

On the opposite, its value is maximum if knowing the value of one of the features provides complete information about the values of all the other ones:

$$TC_{max} = \sum_{i=1}^{n} H(X_i) - \max_{X_j \in F} H(X_j) \tag{11}$$

It is important to note that even though total correlation considers all the redundancies that are present in a set of variables, these redundancies may be distributed through it in a variety of complicated ways [24]. For example, some variables in the set may be totally inter-redundant while others in the set are completely independent. To understand the existent relationships between variables McGill [36] proposed the concept of *interaction information*, which has been defined a key element in the decomposition of the total correlation [27].

The combination of these two concepts (the joint entropy and total correlation) results in the subset selection function $\Psi(S)$, whose maximization results in the selection of highly informative subsets with low inter-redundancies.

$$max\ \Psi(S), \quad \Psi(S) = H(S) - TC(S) \tag{12}$$

To obtain a measure that lies in a fixed range, equation (12) can be normalized, facilitating its interpretation and comparison across different subsets. This normalization is achieved by dividing each component by its maximum value, resulting in a restriction of the equation's range to $[-1, 1]$

$$\Psi_N(S) = \frac{H(S)}{H(F)} - \frac{TC(S)}{TC_{max}(S)} \tag{13}$$

The $\Psi_N(\cdot)$ operator can be easily transformed to the $[0, 1]$ range. By doing this transformation, we fully define the relevancy constraint of our facet determination criterion:

$$max\ RL(S), \quad RL(S) = \frac{1 + \Psi_N(S)}{2} \tag{14}$$

3.1.2 Minimum redundancy

Despite how relevant are the selected subsets of attributes, to achieve our aim of finding novel partitions we need to search for quality facets that are distinct from currently chosen ones. In this context, distinctness is equal to independence, where we assume that each of these facets should be as independent as possible from each other. Given this assumption, we propose to measure facet dissimilarity using mutual information, resulting in the $\Phi(\cdot)$ function, whose minimization ensues the selection of diverse subsets.

$$min\ \Phi(S), \quad \Phi(S) = \frac{1}{t} \sum_{i=1}^{t} I(S; S_i) \tag{15}$$

However, given that two partitions can have different number of features, normalization is required for our measure to be scaled. Normalization has shown to improve the sensitiveness of the MI, compensating for the MI bias toward multivalued features [46, 35]. For these reason, we define the redundancy constraint of our facet determination the following way:

$$min\ RD(S), \quad RD(S) = \frac{1}{t} \sum_{i=1}^{t} NI(S; S_i) \tag{16}$$

While there are several normalized variants of the MI [35], all of them are bounded to $[0, 1]$, equalling 1 when the two of them are identical and 0 when they are independent. For our criterion we define the normalized MI between two subsets S_A and S_B as the MI divided by their minimum entropy.

$$NI(S_A; S_B) = \frac{I(S_A; S_B)}{min\{H(S_A), H(S_B)\}} \tag{17}$$

3.1.3 A unifying criterion for facet determination

There are several ways to combine these two constraints to generate a normalized criterion. For example, we could take inspiration on the Correlation-based feature selection (CFS) [23] and combine them by dividing on each other. However, we prefer to follow Peng's [40] inspiration and combine them by subtracting one equation from the other. The combination of these two constraints generates a normalized criterion that selects interesting and novel facets in an unsupervised way. We could follow the inspiration on the Correlation-based feature selection (CFS) method [23] and combine them by dividing on each other, however, we believe that both the relevancy and redundancy functions are better suited to subtraction, given their normalized nature, which inspires on the works of Peng et al. [40] and Estevez et al. [19]. For this reason, we define the "unsupervised-maximum-relevancy-minimum-redundancy" (UMRMR) criterion, represented by the $\Omega(\cdot)$ operator, in the following manner:

$$max\ \Omega(S),\quad \Omega(S) = RL(S) - RD(S) \tag{18}$$

The $\Omega(\cdot)$ function is bounded to the [-1,1] interval. However it can be easily be transformed into the [0,1] by simply applying the same approach as in the equation (14).

3.2 Bayesian network factorization

Both our relevancy and redundancy functions make use of joint probability distributions (JPDs) in their calculations, which produces representational, computational and statistical wise problems. These problems are caused by the exponential increase in the number of JPD's parameters with respect to its number of variables. For example, even in the case of a JPD composed of n binary variables, it would need to define and store $2^n - 1$ parameters. These issues are the main barrier to the adoption of the joint entropy and joint MI in both facet determination and unsupervised feature selection procedures, which are produced by the JPD's assumption that all its variables are dependent on each other. To solve these problems we need to factorize this JPD into a more natural and compact representation. The simplest way to factorize this distribution resides in the application of the chain rule, which transforms the JPD into a product of conditional probabilities.

Given a feature set $X = \{X_1, ..., X_n\}$, and a JPD $P(X_1, ..., X_n)$, the probability of each of its members using the chain rule and a topological order can be calculated the following way:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_1, ..., X_{i-1}) \tag{19}$$

However, although this conditional representation is more natural and understandable than the explicit representation of the joint, it is not more compact as it maintains the same number of parameters. To overcome the JPD's exponential size problem, it is necessary to exploit the conditional independences that are present in the data. In this regard, Bayesian networks [39, 28] are able to accomplish this by taking advantage of the local Markov property.

A Bayesian network $B(G, \Theta)$ for a vector of variables X is defined by:

- A directed acyclic graph G that comprise the structure of B and express a set of conditional independencies between its variables.
- A set of local parameters Θ representing the conditional probability distributions of each variable given their parents according to G.

B is a Bayesian network with respect to G if it satisfies the local Markov property: each variable X_i is conditionally independent of its non-descendants given its parent variables. By applying this

property, the equation (19) is transformed into the following one, where each feature F_i is embodied by a variable X_i whose parents according to G are $Pa_G(X_i)$:

$$p(X) = p(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pa_G(X_i)) \tag{20}$$

This factorization allows a tractable evaluation of our UMRMR criterion while considering the dependencies present in data. However, to take advantage of this factorization, a BN model has to be learned for each of the JPDs that are present in the equation (18). Three general approaches to the BN learning problem can be distinguished, each of them adopting a different strategy to the structure search [28]: *score-based learning*, *constraint-based learning* and *Bayesian model averaging*. While the three of them serve the purpose of generating a good solution, we need to choose one whose learning time doesn't overshadow its factorization benefits. This problem is analysed in section 4 with our optimization procedure.

4 Optimization

Our facet determination process can be viewed as a combinatorial optimization problem where our defined UMRMR criterion corresponds to its single-objective function [9]. Combinatorial optimization problems are solved by identifying an optimal solution from a finite group of them. Although an exhaustive search can be conceivably performed if the number of features is not too large, the search becomes quickly intractable as its number increases. For this reason, we choose to apply one type of metaheurisitics that has proven to be adequate for this kind of problems [7]: evolutionary algorithms. Evolutionary algorithms (EAs) are optimisation techniques inspired from natural evolution processes, such as reproduction, mutation, recombination and selection. They are formed of two components: (1) a collection of candidate solutions to the optimization problem that act as individuals of a population, and (2) a fitness function that determines the quality of the solutions with respect to the criterion being optimized.

Given their population-based nature, EAs are suited to the application of dynamic programming techniques. Dynamic programming proposes to divide a problem into a set of smaller ones whose solutions can then be reused. This idea perfectly blends in EAs because the evaluation of an indvidual can be accelerated by having stored some of the components of its evaluation. This idea synergies with one of the BN learning approaches that we meantioned in section 3.2: constraint-based learning.

Constraint-based learning methods try to find a BN model (or more specifically a set of equivalent models) that best explains the independencies present in the data. Two phases can be distinguished in the structure learning process of these algorithms. The first one constructs and undirected graph that represents the *skeleton* of the DAG, while the second transforms it in a potentially directed graph (PDAG). A PDAG represents potential edge orientations for the set of equivalent models. If the edge is directed, then all the members of the equivalence class agree on the orientation of the edge. One the other hand, if it is undirected, there are two DAGs in the equivalence class that disagree on its orientation.

The generation of this structure is achieved by carrying out a series of conditional independence tests, following Pearl's assumption that graphical separation and probabilistic independence imply each other. To answer queries about conditional independencies between a set of variables we rely on doing hypothesis testing, where the null hypothesis states they are conditionally independent. To accept or reject this null hypothesis, it is necessary to define a deviance measure f_D and a significance level. From all the available deviance measures, the χ^2 statistic and the MI are the most commonly

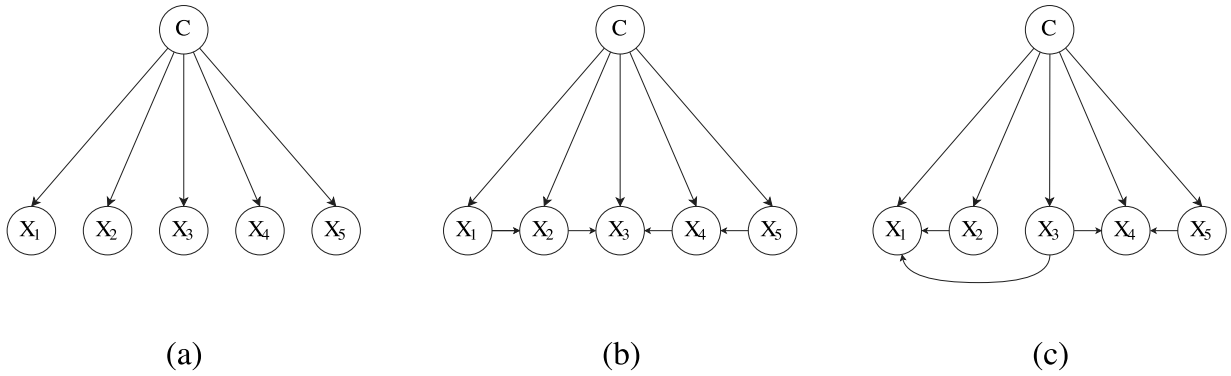

Figure 2: *(a) Naïve Bayes; (b) Tree augmented Naïve Bayes; (c) k-Dependence Bayesian classifier.*

used with discrete data. The typical signifance level is 95 percent, which means that we reject the null hypothesis if the deviance in the observed data has a p-value of 0.05 or less.

The combination of constraint-based learning with EAs in a dynamic programming environment greatly reduces the optimization procedure's complexity, given that each independence test needed to construct a new BN model, will be done just once.

5 Generating alternative partitions from facets

Our facet determination process selects interesting and non-redudant subsets of attributes that are the basis for alternative clustering solutions. Although this process support any clustering algorithm, we propose a model-based clustering algorithm that takes full advantage of the BN model that are associated to each of the facets.

Model-based clustering algorithms partitions the data $\{x_1, ..., x_n\}$ into k different groups, where each of these groups is represented by an univariate distribution. this way, each data object x_i belongs to each cluster c_i to a certain degree. From this, it is sufficient to introduce a hidden node representing the cluster variable $C = \{c_1, ..., c_k\}$ that models a mixture distribution of its components. The cluster variable C is normally assumed to be the parent of all the data attributes $\{X_1, ..., X_n\}$, and thus $P(X_1, ..., X_n)$ can be obtained by combining the probability distributions of the cluster variable, i.e $P(C)$, and the probability distribution of the features given the cluster variable, $P(X_1, ..., X_n|C)$:

$$P(X_1, ..., X_n) = \sum_{i=1}^{k} P(C_i)P(X_1, ..., X_n|C_i) \tag{21}$$

However, this calculation may be inconvenient in the presence of a large number of attributes. Bayesian networks present an intuitive and natural way for performing this type of clustering by reducing the number of parameters to be measured. By applying equation (20), we can produce the following factorization:

$$P(X_1, ..., X_n) = \sum_{i=1}^{k} P(C_i) \prod_{j=1}^{n} P(X_j|Pa_G(X_j), C_i) \tag{22}$$

This equation assumes a Bayesian classifier structure with a latent class variable. There are several possibilites that vary in their faithfulness representing the conditional independencies present in the data. The simplest model is the Naive Bayes, which assumes conditional independency between the

feature variables given the class, and the most complex model the k-dependence Bayesian classifier (k-DB), which allows each feature variable to have a maximum of k parent variables apart from the class variable [4].

The learning process of a BN for clustering is composed of three aspects: (1) determining the cardinality of the variable C, (2) finding the arcs between the attribute variables and (3) estimating the model parameters. While the parameters are normally estimated using the EM algorithm [18], the structure is usually determined by a search process that compares alternatives using a scoring metric. The most commonly used scoring metric with the EM algorithm is the Bayesian information criterion (BIC) [43], which can be defined the following way for a BN model m:

$$BIC(m|D) = logP(D|m, \Theta^*) - \frac{d(m)}{2} log(N) \tag{23}$$

where Θ^* is the maximum likelihood estimate of the parameters, $d(m)$ represents the number of parameters in m and N is the number of data instances. There are other log-likelihood scoring functions like the AIC [1] or the Cheesman-Stutz [10], but we choose to work with BIC because it is the most frequently used score in the literature.

5.1 Searching for optimal partitions

We have developed a hill-climbing algorithm that searches for optimal BN models using the BIC score. The search space of this algorithm is constrained to all the k-DB models that contain the facet's variables. It starts with a k-DB model that has minimal cardinality on the cluster variable and the same feature arcs as the facet factorization network. To explore this space of possible models, 5 operators are defined. A state introduction (SI) operator creates a new model by adding a state to the cluster variable. A state deletion (SD) operator produces the opposite result by removing a state of the cluster variable. Apart from these two, we have three operators that create new k-DB models by modifying the arcs between the BN's feature nodes. The first one is the arc introduction (AI) operator, which, if possible, adds the highest scoring arc between features. An arc deletion (AD) operator is also present, which generates the opposite model of AI, and finally the arc reversal (AR) operator, which returns the highest score model with a reversed feature arc.

6 Conclusions

In this paper, we have introduced a novel methodology for alternative clustering based on a facet determination process. This process is approached as an optimization procedure where the objective function corresponds to our new unsupervised-maximum-relevancy-minimum-redundancy criterion (UMRMR). UMRMR is an information-theoretic criterion that selects feature subsets that are both interesting for the user and different from each other. By taking advantage of the Bayesian network factorization, it is able to compute faithful estimations of the joint entropy and the joint MI that are necessary in its calculation. The optimization process of this criterion returns a set of distinguishing facets that are the basis for our alternative clustering solutions. Once a set of relevant views o data have been chosen, a model-based clustering algorithm is applied to each of them, generating a new partition ready to be studied. This clustering algorithm is also based on Bayesian networks, and thus it is able to reuse the factorizations that were learned through the facet determination process to produce more accurate clustering models.

Acknowledgements

This work has been supported by the Fundación BBVA grants to Scientific Research Teams in Big Data 2016.

References

[1] Akaike H (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22(1), 203-217.

[2] Bae E, Bailey J (2006). Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. *Sixth IEEE International Conference on Data Mining*, 53-62.

[3] Bickel S, Scheffer T (2004). Multi-View Clustering. *The IEEE International Conference on Data Mining series (ICDM)*, 19-26.

[4] Bielza C, Larrañaga P (2014). Discrete Bayesian network classifiers: a survey. *ACM Computing Surveys (CSUR)*, 47(1), 1-43.

[5] Blum C, Roli A (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys (CSUR)*, 35(3), 268-308.

[6] Cai D, Zhang C, He X (2010). Unsupervised feature selection for multi-cluster data. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 333-342.

[7] Calégari P, Coray G, Hertz A, Kobler D, Kuonen P (1999). A taxonomy of evolutionary algorithms in combinatorial optimization. *Journal of Heuristics*, 5(2), 145-158.

[8] Caruana R, Elhawary M, Nguyen N, Smith C (2006). Meta clustering. *Sixth IEEE International Conference on Data Mining*, 107-118.

[9] Charikar M, Guruswami V, Kumar R, Rajagopalan S, Shai A (2000). Combinatorial feature selection problems. 41st IEEE Symposium on Foundations of Computer Science (FOCS), 631-640.

[10] Cheeseman P, Stutz J (1995). Bayesian classifcation (AutoClass): Theory and results. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*. AAAI Press.

[11] Chen T, Zhang NL, Liu T, Poon KM, Wang Y (2012). Model-based multidimensional clustering of categorical data. Artificial Intelligence, 176(1), 2246-2269

[12] Cui Y, Fern XZ, Dy JG (2007). Non-redundant multi-view clustering via orthogonalization. *Seventh IEEE International Conference on Data Mining*, 133-142.

[13] Cui Y, Fern XZ, Dy JG (2010). Learning multiple nonredundant clusterings. *ACM Transactions on Knowledge Discovery from Data*, 4(3), 15.

[14] Dang XH, Bailey J (2010). A hierarchical information theoretic technique for the discovery of non linear alternative clusterings. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 573-582.

[15] Dang XH, Bailey J (2010). Generation of alternative clusterings using the cami approach. *Proceedings of the 2010 SIAM International Conference on Data Mining*, 118-129.

[16] Dasgupta S, Ng V (2010). Mining clustering dimensions. *Proceedings of the 27th International Conference on Machine Learning (ICML-2010)*, 263-270.

[17] Davidson I, Qi Z (2008). Finding alternative clusterings using constraints. *Eighth IEEE International Conference on Data Mining*, 773.778.

[18] Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1), 1-22.

[19] Estévez PA, Tesmer M, Pérez CA, Zurada JM (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2). 189-201

[20] Feng J, Jiao L, Liu F, Sun T, Zhang X (2016). Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images. *Pattern Recognition*, 51, 295-309.

[21] Gondek D, Hogmann T (2007)- Non-redundant data clustering. *Knowledge and Information Systems*, 12(1), 1-24.

[22] Guan Y, Dy JG, Niu D, Ghahramani Z (2010). Variational inference for nonparametric multiple clustering. *MultiClust workshop at KDD*.

[23] Hall MA (1999). Correlation-based feature selection for machine learning.

[24] Han TS (1980). Multiple mutual informations and multiple interactions in frequency data. *Information Control*, 46(1), 26-45.

[25] Handl J, Knowles J, Kell DB (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15), 3201-3212.

[26] Jain P, Meka R, Dhillon IS (2008). Simultaneous unsupervised learning of disparate clusterings. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 1(3), 195-210.

[27] Jakulin A, Bratko I (2003). Quantifying and visualizing attribute interactions: An approach based on entropy. *Journal of Machine Learning Research*.

[28] Koller D, Friedman N (2009). *Probabilistic graphical models: principles and techniques*. The MIT press.

[29] Kriegel HP, Kröger P, Zimek A (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1), 1.

[30] Kriegel HP, Zimek A (2010). Ensemble clustering, alternative clustering, multiview clustering: What can we learn from each other. *MultiClust workshop at KDD*.

[31] Kvalseth TO (1987). Entropy and correlation: Some comments. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(3), 517-519.

[32] Law MH, Topchy AP, Jain AK (2004). Multiobjective data clustering. Proceedings of the 2004 IEEE Computer Society Conference on computer Vision and Pattern Recognition (CVPR-2004), 424-430.

[33] Lazarfeld PF, Henry NW (1968). *Latent Structure Analysis*. Houghton Muffin Company

[34] Liu TF, Zhang NL, Chen P, Liu AH, Poon LK, Wang Y (2015). Greedy learning of latent tree models for multidimensional clustering. Machine Learning, 98(1-2), 301-330

[35] Luo P, Xiong H, Zhan G, Wu J, Shi Z (2009). Information-theoretic distance measures for clustering validation: Generalization and normalization. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1249-1262.

[36] McGill W (1954). Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4), 93-111.

[37] Niu D, Dy J, Ghhramani Z (2012). A nonparametric Bayesian model for multiple clustering with overlapping feature views. *Artificial Intelligence and Statistics*, 814-822.

[38] Niu D, Dy, JG, Jordan MI (2014). Iterative discovery of multiple alternative clustering views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1340-1353.

[39] Pearl J (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, California: Morgan Kaufmann.

[40] Peng H, Long F, Ding C (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8), 1226-1238.

[41] Poom LK, Zhang NL, Liu T, Liu AH (2013). Model-based clustering of high-dimensional data: Variable selection versus facet determination. *International Journal of Approximate Reasoning*, 54(1), 196-215.

[42] Qi Z, Davidson I (2009). A principled and flexible framework for finding alternative clusterings. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 717-726

[43] Schwarz G (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464

[44] Shannon (1948). A mathematical theory of communication. The Bell system Technical Journal, 27, 379-423; 623-656.

[45] Spirtes P, Glymour C (1991). An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review, 9, 62-72.

[46] Strehl A, Ghosh, J (2003). Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3, 583-617

[47] Watanabe S (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1), 66-82.

[48] Zhang NL (2004). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5, 697-723

[49] Zhao B, Kwok JT, Zhang C (2009). Multiple kernel clustering. *Proceedings of the 2009 SIAM International Conference on Data Mining*, 638-649