Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Descriptive Clustering: ILP and CP Formulations with Applications

Thi-Bich-Hanh Dao[1], Chia-Tung Kuo[2], S. S. Ravi[34], Christel Vrain[1], Ian Davidson[2]
[1] LIFO, University of Orléans, France
[2] University of California, Davis
[3] Virginia Tech
[4] University at Albany – SUNY
thi-bich-hanh.dao@univ-orleans.fr, tomkuo@ucdavis.edu, ssravi0@gmail.com,
christel.vrain@univ-orleans.fr, davidson@cs.ucdavis.edu

Abstract

In many settings just finding a good clustering is insufficient and an explanation of the clustering is required. If the features used to perform the clustering are interpretable then methods such as conceptual clustering can be used. However, in many applications this is not the case particularly for image, graph and other complex data. Here we explore the setting where a set of interpretable discrete tags for each instance is available. We formulate the **descriptive clustering** problem as a bi-objective optimization to simultaneously find compact clusters using the features and to describe them using the tags. We present our formulation in a declarative platform and show it can be integrated into a standard iterative algorithm to find all Pareto optimal solutions to the two objectives. Preliminary results demonstrate the utility of our approach on real data sets for images and electronic health care records and that it outperforms single objective and multi-view clustering baselines.

1 Introduction

Clustering is a core unsupervised learning task that aims to partition data into groups to gain a better understanding of the data. However in many domains interpreting the clustering is important. Methods such as conceptual clustering address this by attempting to find descriptions of the clusters using the very features the data was clustered on. However, in many domains that involve images, graphs and other complex data this is not useful. Consider trying to explain an image segmentation using SIFT features or a social network segmentation using the edge structure. In this paper, we explore the scenario where one wants to cluster a data set using a set of features (or a distance matrix) but also form useful descriptions using a set of discrete tags. For example we may wish to not just find communities in a social network but efficiently describe each community by the tags on the people within it. To achieve interpretability, practitioners often perform the clustering and then try to profile each cluster with the tags to describe it. This cluster (try) then describe (test) approach is common in clustering and classification and is typically iterative [Ye and Li, 2002; Cai *et al.*, 2007]. However, since the tags are not directly part of the optimization, the results may not be interpretable.

Here we propose an alternative approach where we *simultaneously* look for clusters which are both useful/compact in one modality (e.g. SIFT features or graph distance) and descriptive in another (e.g. tags). We formulate this *descriptive clustering*[1] problem as a bi-objective optimization problem in a declarative platform and present an algorithm to find the Pareto front of solutions. This work is inherently different from other forms of clustering as discussed in the related work section. In particular, it is not multi-view clustering as the objectives applied to each part of the data are fundamentally different and need not be compatible. It is not conceptual clustering where the concepts are formed from the features used to perform the clustering.

Our contributions can be summarized as follows.

- We describe a bi-criteria descriptive clustering problem formulation which to our knowledge is novel.
- We formulate our problems as an Integer Linear Programming (ILP) problems and show a significantly more efficient Constraint Programming (CP) formulation.
- We show how a standard iterative scheme to find the Pareto optimal solutions can be used to optimize our two objectives simultaneously.
- We demonstrate the usefulness of our approach on real data sets including images and health care records. Our Pareto formulation presents the user a range of clusterings and their descriptions. These are compared against single objective formulations such as multi-view clustering which assumes compatible objectives.

2 Related Work and Novelty

We first discuss superficially similar work and then discuss more closely related work. Approaches such as bi-clustering

[1] Other formulations use the term *descriptive clustering* [Weiss, 2006] but these are for settings where the single set of features naturally form descriptions such as in text.

simultaneously cluster instances and features while our work directly clusters two distinct descriptions of the same instances. Though there has been work on mining discrete patterns using CP [De Raedt *et al.*, 2008] and on distance based clustering using CP [Dao *et al.*, 2017], to our knowledge, there has been no work combining the topics of simultaneously finding clusters on set of features whilst also finding patterns amongst discrete features. Multi-view clustering typically assumes that the same objective is applied to the different views of the data and typically assumes compatibility in that optimizing one objective helps in optimizing the other. Our work makes no such assumptions.

Multi-objective optimization has been studied in the context of clustering. Some earlier work [Runkler, 2007] noted that different clustering objectives and validity indices are not usually consistent and thus proposed to optimize multiple objectives simultaneously. Similar ideas have also been used in the bioinformatics domain to help identify co-expressed genes [Maulik *et al.*, 2009]. Pareto optimization has been used in multiview clustering [Wang *et al.*, 2013] and in combining clustering and classification [Cai *et al.*, 2010]. Some work on multi-objective clustering integrates criteria that represent the partition quality such as compactness and connectivity [Delattre and Hansen, 1980; Handl and Knowles, 2005; Faceli *et al.*, 2007]. Our work is different from all above in that we assume a setting that handles continuous features and categorical tags distinctively and our second objective aims to find meaningful descriptions of clusters.

Predictive clustering is a method of performing classification which aims to find clusters in the input attributes and homogeneity in the class labels at the same time. Earlier work [Langley, 1996] viewed decision trees as a predictive clustering where each leaf is a "cluster" with a homogeneous class label and some attributes (those on its path). More recent work [Ženko *et al.*, 2005] proposed learning predictive clustering rules. In comparison, our work is focused not towards prediction but explanation. In our setting, we have no initial class labels.

Distance-based clustering aims at finding homogeneous clusters only based on a dissimilarity measure between objects. Different declarative frameworks have been developed, which rely on SAT [Davidson *et al.*, 2010], CP [Dao *et al.*, 2013; Dao *et al.*, 2016], ILP [Mueller and Kramer, 2010; Babaki *et al.*, 2014] or quadratic programming [Wang *et al.*, 2014]. Our work uses some of the objectives of these works but only for our first objective function.

Conceptual clustering [Gennari *et al.*, 1989; Fisher, 1987] tries to put objects into classes where each class is defined by a concept expressed in a given description language. The **same set of features** are used to form and describe the clusters. Declarative approaches have been developed using CP [Guns *et al.*, 2013], SAT [Métivier *et al.*, 2012] or ILP [Ouali *et al.*, 2016]. A CP framework for multi-objective conceptual clustering has been developed where the objectives are defined on the concepts [Chabert and Solnon, 2017]. Our work is different in that we study the setting where each instance has **both** continuous features (to form the clusters) and binary tags (to describe them).

3 Descriptive Clustering Formulation

A descriptive clustering problem aims at simultaneously finding compact and descriptive clusters. We define below the compactness and descriptiveness requirements. Let X (the feature matrix) denote a $n \times f$ matrix of n data instances with f continuous features and let X_i be the i-th row of X. Let D (the descriptor matrix) be another $n \times r$ *boolean* matrix of the same n instances, each with r tag indicators. For example, D can represent some one-hot encoded categorical features; e.g. if the j-th column encodes $\texttt{gender} = \texttt{male}$, then $D_{ij} = 1$ indicates that the i-th instance is a male.

3.1 Variables

Cluster Indication Matrix Z: an $(n \times k)$ boolean matrix. Its entries serve as the cluster indicators, $Z_{ic} = 1$ indicates the i-th instance is in the c-th cluster. Let Z_i be the i-th row of Z.
Cluster Description Matrix S: a $(k \times r)$ auxiliary boolean matrix. $S_{cp} = 1$ means the p-th tag is included in the description for the c-th cluster.

3.2 Constraints

Here we outline the constraints to ensure that the matrices Z and S match the requirements of defining a set partition and a useful set of descriptors, respectively.

Partitioning constraints: They enforce valid clustering and descriptions.

$$
\begin{array}{lll}
\text{(C1)} & \forall i = 1, \ldots, n, & \sum_{c=1}^{k} Z_{ic} = 1 \\
\text{(C2)} & \forall c = 1, \ldots, k, & \sum_{i=1}^{n} Z_{ic} \geq 1 \\
\text{(C3)} & Z_{11} = 1 & \\
\text{(C4)} & \forall i = 2, \ldots, n, \forall c = 2, \ldots, k, & \sum_{j=1}^{i-1} Z_{jc-1} \geq Z_{ic} \\
\text{(C5)} & \forall c = 1, \ldots, k, & \sum_{p=1}^{r} S_{cp} \geq 1
\end{array}
$$

(C1) enforces a valid clustering: each instance is in exactly one cluster. (C2) enforces non-empty clusters: each cluster has at least one instance. (C3) and (C4) break symmetries among clusterings: the first instance is in the first cluster and if instance i is in cluster c then cluster $c-1$ must have an instance j such that $j < i$. (C5) enforces a valid non-empty description: each cluster description contains at least one tag.

Description of the clusters: Each cluster is described by a subset of tags. We introduce two integer variables α and β. The constraints below enforce the link between cluster composition and description.

$$
\begin{array}{ll}
\text{(C6)} & \forall c = 1, \ldots, k, \ \forall i = 1, \ldots, n, \\
& Z_{ic} = 1 \implies \sum_{p=1}^{r} S_{cp}(1 - D_{ip}) \leq \alpha \\
\text{(C7)} & \forall c = 1, \ldots, k, \ \forall p = 1, \ldots, r, \\
& S_{cp} = 1 \iff \sum_{i=1}^{n} Z_{ic}(1 - D_{ip}) \leq \beta
\end{array}
$$

(C6) states that if an instance is in a cluster then it satisfies *most* of its description (up to α exceptions). (C7) demands that a tag is included in a cluster's description if and only if *most* of the instances in the cluster (up to β exceptions) possess the tag.

These constraints tolerate disagreement and are useful for datasets with very sparse tags. For datasets with more dense

tags, a stronger version is as follows:

$$\begin{array}{ll}\text{(C6')} & \forall c = 1,\ldots,k,\ \forall i = 1,\ldots,n,\\ & Z_{ic} = 1 \implies \sum_{p=1}^{r} S_{cp}(1 - D_{ip}) = 0\\ \text{(C7')} & \forall c = 1,\ldots,k,\ \forall p = 1,\ldots,r,\\ & S_{cp} = 1 \iff \sum_{i=1}^{n} Z_{ic}(1 - D_{ip}) = 0\end{array}$$

(C6') ensures that each instance must satisfy all the tags that describe the cluster containing it. (C7') states that for each cluster, a tag is included in its description iff all instances in the cluster have this tag.

Note that (C6') and (C7') *do not* define conceptual clustering such as defined in [De Raedt *et al.*, 2008]. Indeed in (C6') we only enforce one direction ($\implies$ instead of $\iff$ as in [De Raedt *et al.*, 2008]) since it is possible for an instance to satisfy the descriptions of more than one cluster. In those cases, the instance can be placed in any such cluster.

3.3 Objectives

We divide the objectives into two categories: feature-focused and tag/descriptor-focused, and within each introduce more than one. Our bi-objective framework refers to simultaneously optimizing one objective from each category.

Category #1: Feature-focused (compactness)
This type of objective can be from a range of typical clustering objectives which aim to find compact clusters. The instances are positioned in some metric space with a distance $d(\cdot,\cdot)$ defined over pairs of instances.

- **Diameter**: This is one of the commonly used objectives. The diameter of a cluster is defined to be the maximum distance between any pair of instances in that cluster; and the diameter of the entire clustering is then the maximum of all those diameters. Hence we minimize:
$$f(Z,S) = \max_{i<j,i,j=1}^{n} Z_i Z_j^T d(X_i, X_j)$$
- **Sum of Within-cluster Distances**: Another common objective is to minimize the sum over the distances between all pairs of instances within each cluster:
$$f(Z,S) = \sum_{i<j,i,j=1}^{n} Z_i Z_j^T d(X_i, X_j)$$

Category #2: Descriptor-focused (descriptiveness)
Our aim is to create collections of tags/properties that are useful for describing a cluster. However in practice the tags in some domains are very sparsely or noisily labeled and enforcing constraints that are too strong is not appropriate. We describe three objectives, which characterize varying amounts of relaxation in the requirements and then discuss scenarios where each is most appropriate.

- **Max-Min Complete Tag Agreement (MMCTA):** This is the strongest objective where we look for clusters where each instance in a cluster shares the common tags with *all* other instances in the same cluster. This objective is subject to the constraints (C1)-(C5), (C6') and (C7'). Accordingly, we aim to maximize $g(Z,S) = q$, with q being the size of the smallest tag set for a cluster:
$$q = \min_c\{\sum_{p=1}^{r} S_{cp}\}$$
This is most useful when the tags are well populated and believed to have little noise in them.
- **Minimize Tag α-β Disagreement (MTD):** This objective is similar to the above except that now we allow *some* instances of each cluster to *opt-out* by not sharing the common tags. This could be desirable if we know the tags contain noise such that some instances will have features similar to other instances in the cluster but just not share the same tags. Accordingly, we aim to minimize such violations/disagreements, subject to the constraints (C1)-(C5), (C6) and (C7):
$$g(Z,S,\alpha,\beta) = \alpha + \beta$$
This is useful when the tags are sparse and/or when we believe there is significant noise in the tagging process.
- **Max-Min Neighborhood Agreement (MMNA):** In this case we no longer require instances in a cluster to share the *same* tags; instead, we demand a weaker form of tag sharing. Specifically, within each cluster, every instance must share *at least* some given number of tags, q with another instance in the cluster. Note that it need *not* be the same q tags for different pairs. The objective to be maximized is $g(Z,S,q) = q$ under the constraints (C1)-(C5), (C6'), (C7') and
$$\forall i,j = 1,\ldots,n,\ Z_i Z_j^T = 1 \implies \sum_{p=1}^{r} D_{ip} D_{jp} \geq q$$

4 Two Methods For Descriptive Clustering

Our formulation can be implemented using ILP or CP. Surprisingly, we find that the CP formulation is significantly (orders of magnitude) faster than the ILP formulation.

4.1 An ILP Formulation

All our constraints and objectives introduced above are or can be transformed into linear constraints. Constraint (C6) is equivalent to $\forall c = 1,\ldots,k,\ \forall i = 1,\ldots,n,$

$$\sum_{p=1}^{r}(S_{cp} + Z_{ic} - 1)(1 - D_{ip}) \leq \alpha$$

and (C7) is equivalent to $\forall c = 1,\ldots,k,\ \forall p = 1,\ldots,r,$

$$\begin{array}{l}\sum_{i=1}^{n}(S_{cp} + Z_{ic} - 1)(1 - D_{ip}) \leq \beta\\ (n+1)S_{cp} \geq 1 + \beta - \sum_{i=1}^{n} Z_{ic}(1 - D_{ip})\end{array}$$

The same method can be used to transform the objectives such as diameter, MMCTA or MMNA into linear form. The problem therefore becomes an integer linear program.

4.2 A CP Formulation

In our CP model, besides using the variables Z and S, we introduce n integer variables $G_i \in \{1,..,k\}$ $(1 \leq i \leq n)$: $G_i = c$ means instance i is in cluster c. (C1) is ensured by channeling constraints on Z and G: $Z_{ic} = 1 \iff G_i = c$. (C2), (C3) and (C4) are ensured by the value precedence constraint $precede(G,[1,..,k])$ [Law and Lee, 2004] and $atleast(1,G,k)$ (at least one variable in G has the value k). Global constraints introduced in [Dao *et al.*, 2017] are used to bind the feature-focused objective variable and the variables in G. Constraints (C6)-(C7) and (C6')-(C7') are expressed using reified constraints.

For MMCTA and MMNA problems we observe that when the domain of the descriptor-focused objective variable q is

$[\underline{q}, \overline{q})$, if two instances share less than $\underline{q}$ tags, they cannot be in the same cluster. Exploiting this observation we have developed a global constraint to bind q and G, which enforces the relation:

$$\forall i, j = 1, .., n,\ G_i = G_j \Longrightarrow \sum_{p=1}^{r} D_{ip} D_{jp} \geq q$$

This constraint revises the upper bound $\overline{q}$ of q using the variables G_i that have been assigned, and also propagates $G_i \neq G_j$ for all pairs of instances i, j that share less tags than the lower bound $\underline{q}$ of q. It maintains bound consistency for q and a partial domain consistency for the variables G_i's.

For each objective, we develop a heuristic search strategy such that at each choice point, a variable G_i with the smallest remaining domain is chosen. All values c in the domain of G_i are examined and the one that gives the best improvement on the objective variable is chosen. A restart mechanism is used to switch between the defined search strategy and a standard search strategy based on the degree of the variables.

Our formulation differs from recent approaches using ILP [Ouali *et al.*, 2016] or CP [Chabert and Solnon, 2017] which first preprocess the data using frequent pattern mining to generate formal concepts and then use ILP/CP to form clusters by choosing a subset of concepts. These works are in the setting of traditional conceptual clustering but cannot be used in our case since cluster descriptions are not formal concepts.

5 Simultaneously Finding Compact & Descriptive Clusters

Let f be a feature-focused objective that we aim to minimize and g be a descriptor-focused one that we aim to maximize. These objectives in general are not compatible and the trade-offs between them are the Pareto optimal solutions of our bi-objective problem. A solution x **dominates** another solution x' if $f(x) \leq f(x')$, $g(x) \geq g(x')$ and either $f(x) < f(x')$ or $g(x) > g(x')$. A solution x is **Pareto optimal** if x is not dominated by any other solution. The set of all such tuples form the **Pareto front**. Each point in the Pareto front represents a state among the objectives in which no objective can be improved without jeopardizing another. Thus, we can treat them as being examples of **optimal trade-offs** between optimizing both objectives. From these solutions, a user typically selects one that is most suitable for the intended purpose.

Algorithm 1 presents a general iterative scheme to find a complete and minimal set of Pareto optimal solutions using our earlier defined constraints $\mathcal{C}$ as sub-problems. This is standard scheme for any bi-objective optimization, and it has been used in CP but for different objectives [Dao *et al.*, 2017]. Our claim that Algorithm 1 returns a minimal and complete set of Pareto optimal solutions can be readily established with a similar proof. Bi-objective optimization in one search has been proposed in [Gavanelli, 2002] and used in [Chabert and Solnon, 2017] for conceptual clustering. This method, in one optimization phase, searches for new solutions and for each solution found, dynamically adds constraints to prevent dominated solutions, until no more solutions are found. However, this method is not appropriate in our case since the feature-focused objective has a floating point value, the improvements are usually very small, which makes the method converge very slowly. Algorithm 1 has better performance since it deals only with the best value of each objective.

Algorithm 1: Compute complete Pareto front

Input: Features X, tags D and number of clusters k.
Output: A complete Pareto front $\mathcal{P}$.

$\mathcal{P} \leftarrow \emptyset$;
$s_1^f \leftarrow$ minimize f subject to $\mathcal{C}$;
$i \leftarrow 1$;
while $s_i^f \neq$ *NULL* **do**
  $s_i^g \leftarrow$ maximize g subject to $\mathcal{C} \cup \{f \leq f(s_i^f)\}$;
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{s_i^g\}$;
  $i \leftarrow i + 1$;
  $s_i^f \leftarrow$ minimize f subject to $\mathcal{C} \cup \{g > g(s_{i-1}^g)\}$;
return $\mathcal{P}$;

6 Experiments

Here we aim to empirically demonstrate the usefulness of our approach. In particular we would like to address the following three questions.

- Do our objective functions produce clusters that have meaningful descriptions?
- Does the bi-objective optimization framework produce a non-trivial Pareto front?
- How do our results compare with base-line non-Pareto optimization formulations?

The first and third questions test the usefulness of our objectives in practice whilst the second question addresses the premise that the objectives are naturally not compatible and hence require Pareto optimization. *If the objective functions were compatible, the Pareto front would contain just one solution as optimizing one objective optimizes the other*. ILP models are implemented in Gurobi using its MATLAB interface. CP models are implemented using Gecode solver.

6.1 Clustering Tagged Images

In this experiment, we try to cluster a set of tagged animal images, which were used for attribute-based classification [Lampert *et al.*, 2009]. The data set contains 30000 images from 50 classes of animals and 85 distinct (binary) tags describing the animals such as black, stripes, water, etc. Each class is associated with a (non-empty) subset of the 85 tags. We randomly sample 100 images from each of the first 10 animal classes: `antelope`, `grizzly bear`, `killer whale`, `beaver`, `dalmatian`, `persian cat`, `horse`, `german shepherd`, `blue whale`, `siamese cat`. We cluster the data using pairwise Euclidean distance between images based on the 2000 dimensional SIFT features used in [Lampert *et al.*, 2009] and describe it using the 85 tags.

For this data since the tags are well populated, the two objectives are minimizing maximum cluster diameter (based on the SIFT features) and maximizing minimum neighborhood agreement (based on the tags). We run our bi-objective formulation with $k = 5$ and present its Pareto front in Figure

Cl#	Composition by animals	Description by tags
C1	1 grizzly bear, 2 dalmatian, 1 horse, 2 blue whale	big, fast, strong, muscle, newworld, smart
C2	5 antelope, 2 grizzly bear, 2 beaver, 5 dalmatian, 5 persian cat, 5 horse, 6 german shepherd, 3 siamese cat	furry, chewteeth, fast, quadrapedal, newworld, ground
C3	69 beaver, 64 dalmatian, 42 persian cat, 29 blue whale, 42 siamese cat	tail, fast, newworld, timid, smart, solitary
C4	100 killer whale, 69 blue whale, 1 siamese cat	tail, fast, fish, smart
C5	95 antelope, 97 grizzly bear, 29 beaver, 29 dalmatian, 53 persian cat, 94 horse, 94 german shepherd, 54 siamese cat	furry, chewteeth, fast, quadrapedal, newworld, ground

(a) First Pareto point: Diameter minimized. MMCTA=4. MMNA=11

Cl#	Composition by animals	Description by tags
C1	2 antelope, 4 dalmatian, 2 horse, 3 german shepherd, 4 siamese cat	furry, lean, longleg, tail, chewteeth, walks, fast, muscle, quadrapedal, active, agility, newworld, oldworld, ground
C2	2 beaver, 1 persian cat, 1 horse, 1 german shepherd	furry, tail, chewteeth, fast, quadrapedal, agility, newworld, ground, smart
C3	100 grizzly bear, 98 beaver, 99 persian cat, 1 siamese cat	furry, paws, chewteeth, claws, fast, quadrapedal, fish, newworld, ground, smart, solitary
C4	100 killer whale, 100 blue whale	spots, hairless, toughskin, big, bulbous, flippers, tail, strainteeth, swims, fast, strong, fish, plankton, arctic, ocean, water, smart, group
C5	98 antelope, 96 dalmatian, 97 horse, 96 german shepherd, 95 siamese cat	furry, lean, longleg, tail, chewteeth, walks, fast, muscle, quadrapedal, active, agility, newworld, oldworld, ground

(b) Third Pareto point. MMCTA=9, MMNA=15

Cl#	Composition by animals	Description by tags
C1	100 antelope, 100 dalmatian	furry, big, lean, longleg, tail, chewteeth, walks, fast, strong, muscle, quadrapedal, active, agility, newworld, oldworld, ground, timid, group
C2	100 horse, 99 german shepherd, 98 siamese cat	black, brown, gray, patches, furry, lean, longleg, tail, chewteeth, walks, fast, muscle, quadrapedal, active, agility, newworld, oldworld, ground, smart, domestic
C3	100 grizzly bear, 100 beaver, 1 siamese cat	brown, furry, paws, chewteeth, claws, fast, muscle, quadrapedal, active, nocturnal, fish, newworld, ground, smart, solitary
C4	100 killer whale, 100 blue whale	spots, hairless, toughskin, big, bulbous, flippers, tail, strainteeth, swims, fast, strong, fish, plankton, arctic, ocean, water, smart, group
C5	100 persian cat, 1 german shepherd, 1 siamese cat	gray, furry, pads, paws, tail, chewteeth, meatteeth, claws, walks, fast, quadrapedal, agility, meat, newworld, oldworld, ground, smart, solitary, domestic

(c) Fifth Pareto point: MMNA maximized. MMCTA=15, MMNA=18

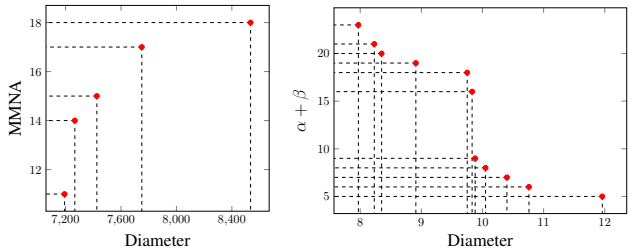
Figure 1: Compositions and descriptions of the chosen clusterings

Cl#	Size	Description
C1	28	Sleepy
C2	25	Light
C3	22	Dizzy
C4	20	Concentrate
C5	14	Sleepy, Light, Dizzy, Concentrate, Vision

Figure 2: Compositions and descriptions of the center point of the Pareto front for our medical data set.

3(a). The clusterings that correspond to some points in the Pareto front are given in Figure 1. If we favor compact clusters to better descriptions (point closest to the origin in Figure 3(a)) we can have clusters described by very few tags, as presented in Figure 1(a). The middle point in the Pareto front, Figure 1(b) trades-off the two objectives and reaches a very intuitive compromise. Long haired persian cats are grouped with similar longer haired animals such as bears, whilst the shorter haired siamese cats are grouped with many shorter haired animals such as dogs and antelopes. The last clustering of the Pareto front favors long description of the clusters, Figure 1(c). The user is free to choose which compromise between the two objectives matches the requirements. *It is important to note that the animal labels are not given to the algorithm.*

The strength of CP resides on search strategies and on its power for pruning inconsistent values. Therefore for MMCTA and MMNA problems where relations can be exploited in our new global constraint, the CP models **always** perform better than ILP models. For instance on our 1000 instance image dataset, our CP model with MMNA takes about 13 seconds to compute the complete Pareto front on a laptop while the ILP model takes more than 6 hours on a 48 core cluster.



(a) Sampled animal images. (b) Personal health record data.

Figure 3: Pareto fronts for the two experiments. Since there are many points, the objectives are not compatible.

6.2 Clustering Personal Health Records

In this smaller experiment we apply our method to a data set of personal health records from the concussion restoration care center[2]. Nevertheless the result is significant as it shows

[2] http://navymedicine.navylive.dodlive.mil/archives/tag/concussion-restoration-care-center

that with even a smaller data set the Pareto front can be relatively large. The data set consists of 109 individuals each with 20 continuous features that are diagnostic scores (i.e. neurological test results) and 14 binary tags that are the symptoms of their condition (Blurred Vision, Fatigue, Confusion etc.). We normalize the continuous features and compute the Euclidean distance between all pairs. We apply our bi-objective formulation (with $k = 5$) where the first objective minimizes the diameter and the second objective looks for minimum tag disagreement (MTD) within a cluster. The choice of the second objective is due to the fact that the tags are very sparse in this data set and the formulation of complete tag agreement would have led to infeasibility immediately.

We present the (approximate) Pareto front in Figure 3(b). It is approximate since we set a time limit and have the solver return the feasible solutions with the best objective up to the time limit. Using domain knowledge, a user may further examine one or more of these solutions. We present the composition and the description of a middle solution in Figure 2 ($\alpha + \beta = 16$). As $\alpha + \beta$ becomes even larger, some clusters contain many tags yet many instances are allowed to disagree with the description.

6.3 Comparison to Baseline non-Pareto Formulations

A natural question is how classic clustering methods such as k-means and multi-view clustering would perform on the setting we describe. Such methods return just a single clustering and it is interesting to explore if it is a clustering on the Pareto front. To apply these methods, we must first carefully combine both the features and tags into one vector and weight the tags more so that the features do not dominate the tags. This was a sensitive trial and error process and the *best* results are shown in Table 1, where the most common tags are given for each cluster. Though the results seem superficially similar to our own method, there are core differences. Firstly, the groupings are contradictory. For example the short haired siamese cats are grouped with bears and the longer haired persian cats were grouped with short haired antelopes: this was never the case with our results for the interior points in the Pareto front. We can also see that the descriptions are contradictory. For example the first cluster describes a cluster of grizzly bears as domesticated and siamese cats as fast, nocturnal and can swim. Similarly for the second cluster, antelopes are described as inactive, solitary and domesticated, while persian cats are described as being fast and active.

Multi-view Clustering. We explored applying multi-view spectral clustering [Zhou and Burges, 2007] with one view being the features and the other view the tags. This work effectively models both sets of descriptors as a graph and the tags are then clustered based on group-wise similarity rather than overlap as our objectives do. Multi-view clustering here assumes both views are **compatible** with each other and combines both objectives into one. Compatible here means that a change that minimizes one objective will also minimize the other. There is no reason this is the case and the benefit of a Pareto formulation is that it finds a trade-off between optimizing both objectives rather than assuming compatibility. We present the results on the image tagged data set (see Table 2).

Cl#	Composition by animals	Ten most common tags
C1	24 german shepherds, 100 grizzly bear, 99 siamese cats	domestic, tail, fast, nocturnal, fish, swim
C2	100 antelopes, 100 persian cats	claws, inactive, solitary, domestic, meat, fast, muscle, quadrapedal, active, agility
C3	76 german shephards, 100 horses, 100 dalmatians	longleg, tail, chewteeth, walks, fast, muscle, quadrapedal, active, agility, newworld, ground
C4	100 blue whales, 100 killer whales, 1 siamese cats	spots, hairless, toughskin, big, bulbous, flippers, tail, strainteeth, water, smart
C5	100 Beavers	brown, furry, tail, chewteeth, fast, muscle, quadrapedal, active, agility, smart

Table 1: Applying k-means clustering to concatenated features and tags vectors for the Tagged Images data set. Note this solution is quite different to the solutions in the Pareto front Figure 1.

Cl#	Composition by animals	Ten most common tags
C1	51 blue whales, 24 german shepherds, 100 grizzly bear, 99 siamese cats	domesticated, fish, flippers, tail, fast, nocturnal, fish, swims
C2	100 antelopes, 33 persian cats, 68 killer whales	claws, strainteeth, inactive, toughskin, domestic, meat, fast, muscle, active, agility
C3	76 german shephards, 100 horses, 100 dalmatians	fast, chewteeth, agile, walks, fast, muscle, quadrapedal, active, furry, newworld, ground
C4	49 blue whales, 42 killer whales, 1 siamese cats	spots, hairless, toughskin, big, bulbous, flippers, tail, strainteeth, water, smart
C5	100 Beavers, 67 persian cats	brown, furry, tail, chewteeth, fast, muscle, quadrapedal, active, agility, smart

Table 2: Applying multi-view spectral clustering to concatenated features and tags vectors for the Tagged Images data set. Note this solution is quite different to the solutions in the Pareto front Figure 1.

We find again the descriptions are unusual when looking at the content. For example cluster 1 describes whales as domesticated and nocturnal and cluster 2 has antelopes having strainteeth, being inactive and domesticated.

7 Conclusions and Future Work

Existing clustering methods focus on clustering a set of points typically modeled as points in a space. Our contribution in this paper is to formulate and explore *descriptive clustering* where each instance is described by a set of features upon which we find a compact set of clusters and simultaneously a set of meaningful tags which describes the clusters. We explore a declarative formulation as a Pareto optimization problem as the two aims need not be compatible (as our experiments show). Since our Pareto fronts do not contain just one point, we indeed find these aims are incompatible and our methods allow users to choose an appropriate trade-off. We explore CP and ILP implementations and observe that CP global constraints allow a much more efficient computation. Baseline k-means and multi-view spectral clustering methods produce a single and very different clustering. Our experiment on health care records shows an interesting result that even small data sets can have a relatively large Pareto front. Future work will explore applications to areas such as social networks and further enhancing the scalability.

References

[Babaki *et al.*, 2014] Behrouz Babaki, Tias Guns, and Siegfried Nijssen. Constrained clustering using column generation. In *CPAIOR*, pages 438–454, 2014.

[Cai *et al.*, 2007] Weiling Cai, Songcan Chen, and Daoqiang Zhang. Robust fuzzy relational classifier incorporating the soft class labels. *Pattern Recognition Letters*, 28:2250–2263, 2007.

[Cai *et al.*, 2010] Weiling Cai, Songcan Chen, and Daoqiang Zhang. A multiobjective simultaneous learning framework for clustering and classification. *IEEE Trans. on Neural Networks*, 21(2):185–200, 2010.

[Chabert and Solnon, 2017] Maxime Chabert and Christine Solnon. Constraint programming for multi-criteria conceptual clustering. In *CP 2017*, pages 460–476, 2017.

[Dao *et al.*, 2013] Thi-Bich-Hanh Dao, Khanh-Chuong Duong, and Christel Vrain. A Declarative Framework for Constrained Clustering. In *ECML/PKDD 2013*, pages 419–434, 2013.

[Dao *et al.*, 2016] Thi-Bich-Hanh Dao, Christel Vrain, Khanh-Chuong Duong, and Ian Davidson. A Framework for Actionable Clustering using Constraint Programming. In *ECAI 2016*, pages 453–461, 2016.

[Dao *et al.*, 2017] Thi-Bich-Hanh Dao, Khanh-Chuong Duong, and Christel Vrain. Constrained clustering by constraint programming. *Artificial Intelligence*, 244:70–94, 2017.

[Davidson *et al.*, 2010] Ian Davidson, SS Ravi, and Leonid Shamis. A SAT-based framework for efficient constrained clustering. In *SDM 2010*, pages 94–105, 2010.

[De Raedt *et al.*, 2008] Luc De Raedt, Tias Guns, and Siegfried Nijssen. Constraint programming for itemset mining. In *KDD 2008*, pages 204–212. ACM, 2008.

[Delattre and Hansen, 1980] Michel Delattre and Pierre Hansen. Bicriterion cluster analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2(4):277–291, 1980.

[Faceli *et al.*, 2007] Katti Faceli, André Carlos Ponce Leon Ferreira de Carvalho, and Marcílio Carlos Pereira de Souto. Multi-objective clustering ensemble. *Int. J. Hybrid Intell. Syst.*, 4(3):145–156, 2007.

[Fisher, 1987] Douglas H Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine learning*, 2(2):139–172, 1987.

[Gavanelli, 2002] Marco Gavanelli. An Algorithm for Multi-Criteria Optimization in CSPs. In Frank van Harmelen, editor, *ECAI 2002*, pages 136–140, 2002.

[Gennari *et al.*, 1989] John H Gennari, Pat Langley, and Doug Fisher. Models of incremental concept formation. *Artificial intelligence*, 40(1-3):11–61, 1989.

[Guns *et al.*, 2013] Tias Guns, Siegfried Nijssen, and Luc De Raedt. k-Pattern set mining under constraints. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):402–418, February 2013.

[Handl and Knowles, 2005] Julia Handl and Joshua D. Knowles. Exploiting the trade-off - the benefits of multiple objectives in data clustering. In *Evolutionary Multi-Criterion Optimization EMO 2005*, pages 547–560, 2005.

[Lampert *et al.*, 2009] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

[Langley, 1996] Pat Langley. *Elements of machine learning*. Morgan Kaufmann, 1996.

[Law and Lee, 2004] Yat Chiu Law and Jimmy Ho-Man Lee. Global constraints for integer and set value precedence. In *CP 2004*, pages 362–376, 2004.

[Maulik *et al.*, 2009] Ujjwal Maulik, Anirban Mukhopadhyay, and Sanghamitra Bandyopadhyay. Combining Pareto-optimal clusters using supervised learning for identifying co-expressed genes. *BMC Bioinformatics*, 10(1):1–16, 2009.

[Métivier *et al.*, 2012] Jean-Philippe Métivier, Patrice Boizumault, Bruno Crémilleux, Mehdi Khiari, and Samir Loudni. Constrained Clustering Using SAT. In *IDA 2012*, pages 207–218, 2012.

[Mueller and Kramer, 2010] Marianne Mueller and Stefan Kramer. Integer Linear Programming Models for Constrained Clustering. In *DS*, pages 159–173, 2010.

[Ouali *et al.*, 2016] A. Ouali, S. Loudni, Y. Lebbah, P. Boizumault, A. Zimmermann, and L. Loukil. Efficiently finding conceptual clustering models with integer linear programming. In *IJCAI'16*, pages 647–654, 2016.

[Runkler, 2007] T. A. Runkler. Pareto optimality of cluster objective and validity functions. In *2007 IEEE International Fuzzy Systems Conference*, pages 1–6, 2007.

[Wang *et al.*, 2013] Xiang Wang, Buyue Qian, Jieping Ye, and Ian Davidson. Multi-objective multi-view spectral clustering via Pareto optimization. In *ICDM*, pages 234–242, 2013.

[Wang *et al.*, 2014] Xiang Wang, Buyue Qian, and Ian Davidson. On constrained spectral clustering and its applications. *Data Min. Knowl. Discov.*, 28(1):1–30, 2014.

[Weiss, 2006] D. Weiss. *Descriptive clustering as a method for exploring text collections*. Doctoral Dissertation, Poznan University of Technology, 2006.

[Ye and Li, 2002] Nong Ye and Xiangyang Li. A scalable, incremental learning algorithm for classification problems. *Computers Industrial Engineering*, 43:677–692, 2002.

[Ženko *et al.*, 2005] Bernard Ženko, Sašo Džeroski, and Jan Struyf. Learning predictive clustering rules. In *KDID*, pages 234–250, 2005.

[Zhou and Burges, 2007] Dengyong Zhou and Christopher JC Burges. Spectral clustering and transductive learning with multiple views. In *ICML*, pages 1159–1166, 2007.