Constrained Minimal Support Set

Javier Larrosa

October 2018

1 Description of the Problem

A large number of typical *data analysis* problems appearing in medicine and in numerous other areas can be formulated in the following way. Consider a *dataset* consisting of two disjoint sets Ω^+ and Ω^- of t-dimensional boolean vectors. For instance, with $t = 8$ we may have,

$$\Omega^+ = \{(0,0,1,0,1,0,0,0),(1,0,1,1,0,0,0,1),(0,1,0,1,0,0,1,1),$$

$$(0,1,1,0,1,1,0,1),(0,0,1,0,1,1,1,1)\}$$

and

$$\Omega^- = \{(1,1,0,0,1,0,1,1),(0,1,0,0,1,1,0,0),(1,0,0,1,1,0,0,1)\}$$

Typically each vector appearing in the dataset corresponds to a patient or a set of similar patients, the vectors in Ω^+ corresponding to patients having a specific medical condition (e.g. pneumonia), while those in Ω^- (the controls in medical language) do not have that condition. The components of the vectors, called *attributes*, or *features*, or sometimes *variables*, represent the results of certain measurements or tests and indicate the presence or absence of certain symptoms (e.g. fever, high blood pressure,...). For instance, if the second attribute in the previous example denotes fever, we can observe that fever is not a necessary symptom for pneumonia since there are patients with fever both in both Ω^+ and Ω^-.

The dataset is assumed to contain sufficient information to characterize the disease. In practice, the dataset usually contains a number of redundant attributes (i.e, not all of them are really needed to characterize the medical

condition). In the following we describe a simple combinatorial optimization problem for eliminating redundant attributes.

Let $T = \{1, 2, \ldots, t\}$ be the set of attributes. A set $S \subseteq T$ is called a *support set* if the projection on S of Ω^+ is disjoint from the projection on S of Ω^-. Recall that the projection of a t-dimensional vector v over $S \subset T$ is a $|S|$-dimensional vector v' obtained removing from v the components not in S. The projection of a set of t-dimensional vectors over $S \subset T$ is the set of $|S|$-dimensional vectors obtained by projecting all the vectors of the original set over S. For instance, if we project Ω^+ over $\{1, 3, 4\}$ we obtain,

$$\{(0, 1, 0), (1, 1, 1), (0, 0, 1), (0, 1, 0), (0, 1, 0)\}$$

Short support sets are important because they are some sort of *short characterizations of the medical condition*. One of the interest of support sets is in diagnosis. Let v be the vector of attributes on the support set of a new patient. If v matches with some vector in the projection on S of Ω^+ it is likely that the patient has the condition. If v matches with some vector in the projection on S of Ω^- it is unlikely that the patient has the condition.

Since we may need to measure the S-attributes of new patients, we may require support sets to satisfy specific constraints. For instance, we may not want S to contain too many expensive tests (for economical reasons), or we may not want S to contain too many aggressive tests (for humanitarian reasons). In this project, for simplicity, we will restrict ourselves to the *At most one* constraint. Let $A \subseteq T$. A support set S satisfies the $AtMostOne(A)$ constraint iff S contains at most one element of A (i.e, $|S \cap A| \leq 1$).

We are now ready to define the problem that we want to consider in the project. Consider two disjuct sets Ω^+ and Ω^- of $t-$dimensional boolean vectors, a natural number k and a list $A_1, A_2, \ldots, A_c \subseteq T$ of sets of attributes. We want to find a set $S \subset T$ such that:

1. S is a support set of Ω^+ and Ω^-
2. its size is bounded by k (that is, $|S| \leq k$).
3. S satisfies the *Atmostone* constraint with respect to $A_1, A_2, \ldots, A_c$.

2 What do you have to do

- A model for the problem with propositional logic in CNF.

- The same model written with MiniZinc[1]. Instances must have the same syntax as the following exemple:

```
       t=8; %number of attributes
k=3; %maximum size of support set
n=5; %number of positive instances
m=3; %number of negative instances
c=4; %number of atMostOne Constraints

omegap=[| 0,0,1,0,1,0,0,0|
          1,0,1,1,0,0,0,1|
          0,1,0,1,0,0,1,1|
          0,1,1,0,1,1,0,1|
          0,0,1,0,1,1,1,1
       |];

omegan=[| 1,1,0,0,1,0,1,1|
          0,1,0,0,1,1,0,0|
          1,0,0,1,1,0,0,1
       |];


atMostOne =
  [{1,2,3},
  {4,5,6},
  {3,5},
  {7,8}];
```

- A random instance generator (in any programming language such as python or C++) to obtain a benchmark with which you can run experiments to see how well MiniZinc performs.

[1]The model must be exactly the same. You can only use boolean variables. Constraints can only be logical operators. The only exception is sum()()=k

3 Documentation

You have to deliver the following items:

- A description of the model in CNF. Describe the model making very clear what are the variables and what do they mean. Also make very clear what is the CNF formula and discuss its assymptotical size in terms of its parameters: $t, |\Omega^+|, ...$
- The MiniZinc model
- A folder containing the instance generator and all the instances that you have used in your experiments
- A brief document reporting an empirical analysis of how well MiniZinc performs on solving this problem.