Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-022-00508-4

REGULAR ARTICLE

On mathematical optimization for clustering categories in contingency tables

Emilio Carrizosa[1] · Vanesa Guerrero[2] · Dolores Romero Morales[3]

Received: 13 October 2021 / Revised: 22 February 2022 / Accepted: 7 June 2022

Abstract

Many applications in data analysis study whether two categorical variables are independent using a function of the entries of their contingency table. Often, the categories of the variables, associated with the rows and columns of the table, are grouped, yielding a less granular representation of the categorical variables. The purpose of this is to attain reasonable sample sizes in the cells of the table and, more importantly, to incorporate expert knowledge on the allowable groupings. However, it is known that the conclusions on independence depend, in general, on the chosen granularity, as in the Simpson paradox. In this paper we propose a methodology to, for a given contingency table and a fixed granularity, find a clustered table with the highest χ^2 statistic. Repeating this procedure for different values of the granularity, we can either identify an *extreme grouping*, namely the largest granularity for which the statistical dependence is still detected, or conclude that it does not exist and that the two variables are dependent regardless of the size of the clustered table. For this problem, we propose an assignment mathematical formulation and a set partitioning one. Our approach is flexible enough to include constraints on the desirable structure of the clusters, such as must-link or cannot-link constraints on the categories that can, or cannot, be merged together, and ensure reasonable sample sizes in the cells of the clustered table from which trustful statistical conclusions can be derived. We illustrate the usefulness of our methodology using a dataset of a medical study.

✉ Vanesa Guerrero
vanesa.guerrero@uc3m.es

Emilio Carrizosa
ecarrizosa@us.es

Dolores Romero Morales
drm.eco@cbs.dk

[1] Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Seville, Spain

[2] Department of Statistics, Universidad Carlos III de Madrid, Getafe, Spain

[3] Department of Economics, Copenhagen Business School, Frederiksberg, Denmark

Published online: 28 June 2022

Keywords Contingency tables · Mathematical optimization · Relational constraints · Clustering

Mathematics Subject Classification 90C90 · 90C27

1 Introduction

Data science comprises a plethora of methods and strategies to extract useful knowledge from raw data. Among other disciplines, such as statistics, computer science or information technology, mathematical optimization plays a crucial role across its tasks (Bottou et al. 2018; Gambella et al. 2021; Olafsson et al. 2008). Much effort has gone into incorporating recent developments in optimization theory and software to tackle data science problems more effectively, such as in regression and classification (Bertsimas and King 2016; Bertsimas and Shioda 2007; Blanquero et al. 2020; Carrizosa and Romero Morales 2013; Carrizosa et al. 2021; Toriello and Vielma 2012), clustering strategies (Benati and García 2014; Carrizosa et al. 2013; Hansen and Jaumard 1997; Hochbaum and Liu 2018; Park et al. 2000; Sağlam et al. 2006), correspondence analysis (van de Velden et al. 2020), dimensionality reduction methods (Carrizosa and Guerrero 2014; Carrizosa et al. 2020; Cunningham and Ghahramani 2015), deep learning (Anderson et al. 2020; Fischetti and Jo 2018) or data visualization (Carrizosa et al. 2017a, 2018a, b, 2019).

There are still many data science problems which do not take advantage of such advancements and ad-hoc strategies are still used, such as, the analysis of the independence of two categorical variables through a function of the entries of their contingency table. Let U and V be two categorical variables, which take on a finite number of values, $u_1, \ldots, u_r$ and $v_1, \ldots, v_c$, respectively. Given a set of n entities for which these variables have been observed, a first summary of their distribution is provided by their *contingency table*, in which the frequency of the event (u_i, v_j), o_{ij}, is collected for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. Table 1 contains an example of a contingency table in which $n = 98$ observations (lowest right corner) are cross-classified according to variable U, which has categories u_1 and u_2 and variable V, which has three categories (v_1, v_2 and v_3). Whereas the inner rectangle in the table contains the joint frequencies o_{ij}, the last row (resp. column) contains the marginal frequencies $o_{.j}$ of V, $j = 1, \ldots, c$ (resp. $o_{i.}$ of U, $i = 1, \ldots, r$).

When the data is cross-classified as in Table 1, the statistical (in)dependence of two categorical variables is usually investigated using the classical χ^2 measure (Pearson 1900; Mirkin 2001), although different approaches exist in the literature (Goodman and

Table 1 Example of a contingency table

	v_1	v_2	v_3	
u_1	7	8	9	24
u_2	8	43	23	74
	15	51	32	98

Kruskal 1979; Joe 1989). The χ^2 coefficient is an estimate of the deviation between the empirical probability distribution of the variables U and V and the probability distribution that we would have if the two variables were statistically independent, and it is given by

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \tag{1}$$

where $e_{ij} = \dfrac{o_{i.}o_{.j}}{n}$. The minimum value of χ^2 in (1) is 0, which occurs if and only if the variables U and V are statistically independent. Therefore, the larger χ^2, the stronger the evidence against independence. The χ^2 statistic in (1) approximates a Chi-squared distribution with $(r-1) \times (c-1)$ degrees of freedom. In Table 1, one gets $\chi^2 = 6.355$, which provides evidence against the statistical independence of the two variables involved ($p-$value$= 0.04$, which tests the null hypothesis of statistical independence) for a 5% significance level. These inferential properties can be derived as long as the observed joint frequencies o_{ij} are large enough for all $i = 1, \ldots, r$ and $j = 1, \ldots, c$. To ensure this, the categories of the variables, associated with the rows and columns of the table, are often grouped, yielding a less granular representation of the categorical variables. Clustering categories in rows and/or columns of a contingency table is also desirable to enhance interpretability and transparency (Baesens et al. 2003; Carrizosa et al. 2017b, 2022; Goodman and Flaxman 2017; Ustun and Rudin 2016), by easing the presentation of the table as well as the conclusions of the analysis from a statistical perspective. Furthermore, constrained clustering allows the analyst to incorporate knowledge about the problem under study and support meaningful decision making (Abin 2019; Śmieja and Wiercioch 2017).

However, it is known that the conclusions on independence depend, in general, on the granularity chosen for each of the categorical variables. For instance, let us consider that variable V in the example above is encoded as $v_1' = v_1$ & v_3 and $v_2' = v_2$. Thus, observations in v_1 and v_2 are now grouped together, yielding the contingency table in Table 2, for which $\chi^2 = 3.519$. Thus, the clustered table has $c' = 2$ columns instead of the $c = 3$ in the initial table and is the one yielding the largest χ^2 among all the tables of that granularity, namely with two columns and two rows. In this case, there is not a significant evidence at a 5% significance level to reject the statistical independence assumption between the variables U and V ($p-$value$= 0.06$). Therefore, the Simpson's paradox (Blyth 1972) arises in this example since the less granular representation of the categorical variables in Table 2 supports a conclusion, namely statistical independence, different from that suggested

Table 2 Contingency table resulting from grouping together the categories v_1 and v_3, yielding category v_1', in Table 1

	v_1'	v_2'	
u_1	16	8	24
u_2	31	43	74
	47	51	98

by the variables before the grouping of categories, namely statistical dependence in Table 1 (Shmueli and Yahav 2017; Tsumoto 2009). Thus, we have identified a so-called *extreme grouping*, namely Table 1 has the largest granularity of the variables U and V for which statistical dependence between them is found.

In this paper we propose a mathematical optimization model to cluster the rows/columns of a contingency table so that the χ^2 statistic is maximized for a fixed granularity of the variables. Solving this problem for different sizes allows us to either identify extreme groupings or conclude that they do not exist. Our approach is flexible enough to include constraints on the desirable structure of the clusters, such as must-link or cannot-link constraints on the categories that can, or cannot, be merged together, and ensure reasonable sample sizes in the cells of the clustered table from which trustful statistical conclusions can be derived. This constrained clustering approach allows us to incorporate background knowledge to support the analysis and extract meaningful conclusions (Abin 2019; Śmieja and Wiercioch 2017).

The problem of clustering the categories of a contingency table to find extreme groupings for a fixed granularity has not been studied as such in the literature. There are however related approaches that involve distributional assumptions or use ad-hoc heuristic procedures that are not flexible enough to include constraints on the clusters. Indeed, Greenacre (1988) proposed a greedy procedure, based on hierarchical clustering, which uses a χ^2 related distance function between row (resp. column) vectors. However, this approach does not guarantee that clustered tables with the highest dependence for a fixed granularity are necessarily found because only a reduced family of allowable groupings, namely a hierarchical structure, are considered. The classical k-means clustering algorithm has been also adapted to the particular case of contingency tables (Govaert 1995; Govaert and Nadif 2007) as well as geometrical approaches, such as the maximum-tangent-plane (Bock 2003), have been developed. Ciampi et al. (2005) propose using the coordinates obtained with correspondence analysis to find a clustering and Álvarez de Toledo et al. (2018) use a similarity measure between the categories to obtain a partition. In order to find homogeneous clusters in document-term matrices, Ailem et al. (2016) propose maximizing a graph modularity criterion and Labiod and Nadif (2011) a community detection one. Whereas these approaches are based on the optimization of a measure of association, some probabilistic approaches have been also studied. In this case, it is assumed that each element of the contingency table is generated according to a probability model, which is tried to be recovered from the data. In this context, Ailem et al. (2017a, b); Riverain and Nadif (2022) propose latent block models to identify a diagonal structure of homogeneous blocks in document-term matrices. Proceeding this way, blocks of zeroes are identified and clustered together, thus yielding joint frequencies in the clustered table which are (close to) zero. A unified framework about the optimization of measures of association and probabilistic approaches is studied by Govaert and Nadif (2010). The aforementioned approaches are unable to deal with the analysis of dependence between variables in sparse tables, namely tables for which some of the observed joint frequencies are equal or close to zero and thus statistical conclusions cannot be inferred. It is well known that the common practice of adding constants to small joint frequencies can disturb the possible statistical dependence structure underlying in a sparse table (Agresti and Yang 1987). If some categories were properly clustered together, this sparsity problem would be removed without damaging the underlying possible relationship between the variables. However, the existing methods cannot incorporate the corresponding constraints to ensure that this removal of sparsity is achieved.

The idea of clustering categories in contingency tables has been also applied to the discretization of continuous variables involved in supervised learning algorithms. To guide the search of the partition, a criterion which assesses the relationship between the intervals in which the continuous variable is split and the target values to be predicted is optimized. For instance, Kerber (1992) uses the χ^2 distance between adjacent intervals to merge them if they are similar enough according to a given threshold. Boulle (2004) proposes a greedy approach and uses the $p-$value associated to the χ^2 statistic of the clustered table to select the discretization. However, these methods fail when constraints have to be imposed to the discretization being sought, such as that each interval in the partition has to have a large enough number of observations or there are rules that have to be accomplished (e.g. minimum or maximum length of the intervals which form the discretization).

In this paper, we propose an assignment and a set partitioning mathematical optimization formulations to cluster rows and/or columns of contingency tables maximizing the χ^2 statistic in (1), as a measure of the strength of the dependence, for a fixed size of the clustered table. Solving this model for different sizes, we can decide whether the statistical dependence can be preserved with the chosen granularity of the variables. If this is the case, we reduce the size of the parameter and solve the maximization problem again. We do this until we find the extreme groupings, or conclude that they do not exist, namely for any size of the reduced table the dependence of the variables can be preserved. Our model can easily be enriched with constraints to incorporate user knowledge on the allowable groups of categories, or to successfully handle sparse tables. With the proposed formulations, even contingency tables as the ones in the numerical section can be tackled using off-the-shelf optimization solvers.

The remainder of the paper is structured as follows. Section 2 states the mathematical optimization model to cluster categories in a contingency table maximizing the χ^2 statistic and imposing structural properties in the clusters. An assignment and a set partitioning formulations for such model are presented in Sect. 3. Finally, Sect. 4 illustrates our methodology and Sect. 5 concludes the paper with some remarks and future research.

2 Problem definition

This section is devoted to presenting a mathematical optimization model to cluster the rows and/or columns of a given contingency table which maximizes the χ^2 statistic in (1) for a fixed granularity of the categorical variables whereas requirements on the clusters, that is conditions about allowable groups of categories or thresholds over the sample sizes in the cells of the clustered table, are also imposed.

Let T_0 be a contingency table representing the counts of outcomes of two categorical variables U and V, which both take a finite set of values (categories), $u_1, \ldots, u_r$ and $v_1, \ldots, v_c$, respectively. Recall that given a sample of n entities, o_{ij} denotes the joint observed frequency of the pair (u_i, v_j), $o_{i.}$ the marginal frequency of u_i and $o_{.j}$ the marginal frequency of v_j, for $i = 1, \dots, r$ and $j = 1, \dots, c$. In order to measure the strength of the association between the variables U and V, the χ^2 statistic as stated in (1) is used. Let $\chi^2(T_0)$ be the value of (1) for data in table T_0.

Given the contingency table T_0, a clustered table T is obtained from it by merging the rows and/or columns of T_0 into a new set of categories (clusters). In other words, a set of k $(k \leq c)$ clusters of columns of T_0, $\{\tilde{v}_1, \dots, \tilde{v}_k\}$, is a partition of the set $\{v_1, \dots, v_c\}$ into k groups such that, for $l, l' = 1, \dots, k$:

- $\tilde{v}_l \subseteq \{v_1, \dots, v_c\}$,
- $\bigcup_{l=1}^{k} \tilde{v}_l = \{v_1, \dots, v_c\}$,
- $\tilde{v}_l \cap \tilde{v}_{l'} = \emptyset, l \neq l'$.

Similarly, row clusters can be also defined as $\tilde{u}_1, \dots, \tilde{u}_s$ $(s \leq r)$. The clustered contingency table T from T_0 has a less granular representation of its categorical variables U and V and has as joint frequencies $\tilde{o}_{ml}$, which are obtained from the sum of the corresponding joint frequencies in T_0, namely $\tilde{o}_{ml} = \sum_{\substack{i:\, u_i \in \tilde{u}_m \\ j:\, v_j \in \tilde{v}_l}} o_{ij}$. In other words, a clustered table T accumulates the corresponding frequencies in T_0. Let $\chi^2(T)$ be the value of (1) in table T.

Clustering the rows and/or columns of a contingency table reduces the value of the χ^2 statistic, that is $\chi^2(T) \leq \chi^2(T_0)$ (see Govaert and Nadif (2018) for a detailed proof). Therefore, to see whether we can preserve the dependence structure between the variables U and V when their categories are clustered, we seek, for a fixed size, the clustered table T which maximizes $\chi^2(T)$. Repeating this procedure for different values of the granularity, we can either identify an extreme grouping or conclude that it does not exist, namely the two variables are dependent regardless of the size of the clustered table. Indeed, in the event of obtaining a clustered table T so that statistical dependence is assumed and its clustered table exhibits independence, then we say that T is an extreme grouping.

In order to obtain new categories in table T, namely the clusters, which are meaningful for the analyst, being able to incorporate prior knowledge about the groups of categories which are allowed or not to be merged would be helpful. In other words, not every possible combination of categories is allowed. Besides easing the interpretability, clustering can be used to deal with sparsity issues in the entries of T_0 looking for aggregations of columns and/or rows which accumulate at least a certain number of observations. Let $\mathcal{T}(T_0)$ be the set of all possible contingency tables resulting from allowable groups of rows and columns of T_0.

The problem of Clustering a Contingency Table (CCT) described above is stated as the following combinatorial optimization problem:

$$\begin{aligned} \max_{T} \quad & \chi^2(T) \\ \text{s.t.} \quad & T \in \mathcal{T}(T_0). \end{aligned} \tag{CCT}$$

(CCT) seeks the table $T \in \mathcal{T}(T_0)$ that maximizes the strength of the association between the variables in the clustered table T measured through the χ^2 statistic in (1), and satisfies the structure imposed on the clusters through the definition of the feasible set $\mathcal{T}(T_0)$. (CCT) is a combinatorial problem for which an assignment (0–1 nonlinear) formulation is proposed in the next section. Other formulations are also possible, such as a set partitioning one, which is stated in Sect. 3.3.

3 An assignment formulation and its set partitioning counterpart

This section is devoted to developing a mathematical optimization formulation for the (CCT) model stated in Sect. 2. An assignment formulation is proposed in Sect. 3.1, in which the decisions to be made are whether a column of the observed contingency table T_0 is assigned to a cluster of categories in the clustered table T or not, yielding a $0-1$ nonlinear optimization model. Section 3.2 is devoted to formally model some structures which could be demanded to the clustered table T to, for instance, get meaningful clusters by incorporating expert knowledge on the allowable groupings or reduce sparsity. Recall that these conditions naturally arise from the problem under study. Finally, a set partitioning formulation is proposed in Sect. 3.3.

Clustering a contingency table T_0 can be done either row-wise (only the rows are clustered while the initial columns in T_0 are maintained), column-wise (only the columns are clustered while the initial rows in T_0 are maintained), or in both directions, this is, column and rows are both clustered into new categories. Whereas our approach is valid for any of these three options, the assignment formulation for (CCT) and its extensions are fully developed column-wise for the sake of clarity.

3.1 The assignment formulation for (CCT) with *k* clusters

Recall that $\{v_1, \dots, v_c\}$ is the set of categories (columns) of variable V in the observed contingency table T_0. The categories in T_0 are aimed to be clustered into k new categories, named as $\tilde{v}_1, \dots, \tilde{v}_k$, $k \leq c$ in such a way that the categories in T form a partition of the ones in T_0 and the χ^2 statistic of the clustered table, that is $\chi^2(T)$, is maximized.

Let y_{jl} for all $j \in \{1, \dots, c\}$ and $l \in \{1, \dots, k\}$ be a binary decision variable defined as

$$y_{jl} = \begin{cases} 1 & \text{if the } j\text{-th category in } T_0(v_j), \text{ is assigned to the } l\text{-th category } (\tilde{v}_l), \text{ in } T, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\chi^2\left(\{y_{jl}\}_{\substack{j\in\{1,\dots,c\}\\ l\in\{1,\dots,k\}}}\right)$ be the χ^2 statistic for the clustered table T, which is defined by the y-variables and (1) as:

$$\chi^2\left(\{y_{jl}\}_{\substack{j\in\{1,\ldots,c\}\\ l\in\{1,\ldots,k\}}}\right)=\sum_{l=1}^{k}\sum_{i=1}^{r}\frac{\left(\sum_{j=1}^{c}(o_{ij}-e_{ij})y_{jl}\right)^2}{\sum_{j=1}^{c}e_{ij}y_{jl}}$$

The problem of clustering the columns of a contingency table T_0 into k clusters maximizing the χ^2 statistic is stated as a 0–1 nonlinear optimization model, which consists of the maximization of a convex function subject to linear constraints, as follows:

$$\begin{aligned}
\max \quad & \chi^2\left(\{y_{jl}\}_{\substack{j\in\{1,\ldots,c\}\\ l\in\{1,\ldots,k\}}}\right) && (2)\\
\text{s.t.} \quad & \sum_{l=1}^{k} y_{jl}=1, \quad j=1,\ldots,c, && (3)\\
& \sum_{j=1}^{c} y_{jl}\geq 1, \quad l=1,\ldots,k, && (4)\\
& y_{jl}\in\{0,1\} \quad j=1,\ldots,c,\ l=1,\ldots,k. && (5)
\end{aligned}$$

Constraint (3) ensures that each category in T_0 goes to just one of the new categories (clusters) in T and constraint (4) imposes that each cluster has at least one category. Finally, constraint (5) defines the binary nature of y-variables. Note that problem (2)–(5) can be enriched with constraints to break the symmetry associated with the clusters.

3.2 Modelling some clustering structures

The assignment formulation in (2)–(5) provides a flexible framework to incorporate additional requirements on the clusters in table T in a straightforward manner, namely as additional linear constraints. We describe in what follows some of the most natural cases, although more complex structures, such as the constrained discretization of continuous variables for supervised learning algorithms, can also be modeled using the y-variables in the assignment formulation stated above.

- **Non-sparsity constraints**: In contingency tables analysis usually happens that some of the observed joint frequencies are equal or close to zero, and thus statistical conclusions cannot be inferred from the distribution of the χ^2 statistic. In order to be able to apply statistical inference theory, sparsity problems in the observed table T_0 might be mitigated by clustering some of its columns by imposing a threshold over the number of observations in each row of the new column (cluster). In other words, the user might require that in each row of the columns in the new table T there are at least β observations. A common value for β is 5. Such condition can be added as a constraint to problem (2)–(5) as follows:

$$\sum_{j=1}^{c} o_{ij} y_{jl} \geq \beta, \quad i = 1, \ldots, r, \; l = 1, \ldots, k. \tag{6}$$

- **Cannot-link constraints**: A cannot-link constraint is used to specify that two or more specified categories in T_0 cannot be associated with the same cluster in T. In its simplest case, namely two categories v_j and $v_{j'}$ in T_0 cannot be grouped together in T, the cannot-link constraint is modeled as

  $$y_{jl} + y_{j'l} \leq 1, \quad l = 1, \ldots, k. \tag{7}$$

  Condition (7) can be easily generalized to accommodate groups of categories in T_0 which cannot belong to the same cluster.

  A complementary set of conditions to cannot-link ones are the so-called must-link constraints, which are used to specify that two or more specified categories in T_0 must be assigned to the same cluster in T. Although these kind of conditions could be also easily modeled in a similar fashion, they can be imposed in a preprocessing step.
- **Relational constraints**: There might be structural conditions among categories which are more complex than the ones given by cannot or must-link constraints. That is the case of, for instance, the existence of a partial order relation $\prec$ between the categories implying that, if two categories belong to one cluster then all the categories *in-between* must belong to the same cluster too. In its simplest case, namely two categories v_j and $v_{j'}$ in T_0 such that v_j precedes $v_{j'}$ in the partial order, the so-called relational constraint is modeled as

  $$y_{jl} + y_{j'l} \leq y_{j''l} + 1, \quad \text{for } v_j \prec v''_j \prec v'_j \text{ and } l = 1, \ldots, k. \tag{8}$$

- **Demand / capacity constraints**: We may also require that each column l (clusters) in T contains at least a_l categories of T_0 and/or no more than b_l, thus establishing demand and/or capacity constraints, respectively, for $l = 1, \ldots, k$. Such conditions can be added as constraints to problem (2)–(5) as follows:

  $$a_l \leq \sum_{j=1}^{c} y_{jl} \leq b_l, \quad l = 1, \ldots, k. \tag{9}$$

- **'et al.' clustering**: Given a contingency table T_0 with c columns, the analyst might be interested in obtaining a clustered table T with k columns in which $k-1$ of its categories are exactly $k-1$ of the categories in T_0 and the k-th category is made up of the aggregation of the remaining $c-k+1$ categories in T_0, that is the *'et al.'* category. This structure is a particular case of (CCT) with k clusters, in which $k-1$ clusters are singletons and the k-th category in the new table comprises $c-(k-1)$

categories. In order to get such structure in T, constraint (4) in the formulation (2)–(5) for (CCT) must be replaced by

$$\sum_{j=1}^{c} y_{jl} = 1, \quad l = 1, \ldots, k-1, \tag{4a}$$

$$\sum_{j=1}^{c} y_{jk} = c - k + 1. \tag{4b}$$

Nevertheless, the number of variables and constraints in the optimization problem defined by (2), (3), (4a), (4b) and (5) can be significantly reduced if the following variables are considered instead:

$$y_j = \begin{cases} 1 & \text{if the } j\text{-th category in } T_0(v_j) \text{ is in } T \text{ as a singleton,} \\ 0 & \text{otherwise.} \end{cases}$$

Using this new definition of y-variables, the χ^2 statistic in (1) is rewritten as

$$\chi^2(\{y_j\}_{j\in\{1,\ldots,c\}}) = \sum_{i=1}^{r} \left\{ \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} y_j + \frac{\left(\sum_{j=1}^{c} (o_{ij} - e_{ij})(1 - y_j)\right)^2}{\sum_{j=1}^{c} e_{ij}(1 - y_j)} \right\}.$$

Therefore, the 0–1 nonlinear formulation for the (CCT) problem with the 'et al.' structure (2), (3), (4a), (4b) and (5) is rewritten as

$$\max \quad \chi^2(\{y_j\}_{j\in\{1,\ldots,c\}}) \tag{10}$$

$$\text{s.t.} \quad \sum_{j=1}^{c} (1 - y_j) = c - k + 1, \tag{11}$$

$$y_j \in \{0, 1\} \quad j = 1, \ldots, c. \tag{12}$$

Constraints (11) and (12) control the number of categories in table T_0 which compose the 'et al.' category in table T and the binary nature of the y-variables, respectively.

3.3 A set partitioning formulation for (CCT)

In this section, an alternative formulation is proposed for (CCT), which assumes that there is a list of permissible aggregations of columns in T_0 that can be used to build the clustered table T. Given such a list, a set partitioning formulation is proposed whose benefits with respect to an assignment one are twofold: first, its continuous relaxation is, in general, tighter (Freling et al. 2003), and second, the so-obtained formulation becomes 0–1 linear.

In order to state a set partitioning formulation for (CCT), let $\{S_1, \dots, S_K\}$ be a family of K subsets of the categories in T_0 given by $\{v_1, \dots, v_c\}$. These subsets represent the list of allowable aggregations of columns in T_0, that is the list of clusters which can be used to build the columns in the clustered table T. Let A be a $c \times K$ 0–1 matrix with entries a_{jp} for all $j \in \{1, \dots, c\}$ and $p \in \{1, \dots, K\}$ defined by

$$a_{jp} = \begin{cases} 1 & \text{if } v_j \in S_p, \\ 0 & \text{otherwise,} \end{cases}$$

and let x_p for all $p \in \{1, \dots, K\}$ be a binary decision variable defined by

$$x_p = \begin{cases} 1 & \text{if } S_p \text{ is a column of the clustered table } T, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\chi^2(\{x_p\}_{p\in\{1,\dots,K\}})$ be the χ^2 statistic for the clustered table stated as

$$\chi^2(\{x_p\}_{p\in\{1,\dots,K\}}) = \sum_{i=1}^{r}\sum_{p=1}^{K} \frac{\left(\sum_{j=1}^{c}(o_{ij} - e_{ij})a_{jp}\right)^2}{\sum_{j=1}^{c} e_{ij}a_{jp}} x_p.$$

Then, the set partitioning formulation for problem (CCT) is stated as:

$$\begin{aligned} \max \quad & \chi^2(\{x_p\}_{p\in\{1,\dots,K\}}) && (13) \\ \text{s.t.} \quad & \sum_{p=1}^{K} a_{jp}x_p = 1, \quad j = 1, \dots, c && (14) \\ & x_p \in \{0, 1\} \quad k = 1, \dots, K. && (15) \end{aligned}$$

Whereas (14) ensures that each column of T_0 belongs to just one cluster in T, constraint (15) imposes the binary nature of the x-variables. We point out that (13)–(15) is a 0–1 linear optimization problem, which can easily accommodate the structures discussed in Sect. 3.2. For instance, the relational condition requiring that if columns v_j and $v_{j'}$ in T_0 belong to the same cluster in table T then column $v_{j''}$ belongs also to that cluster is defined through one of the S-sets as $\{v_j, v_{j'}, v_{j''}\}$.

4 Illustrative examples

In order to illustrate the methodology proposed in this paper, a contingency table T_0 obtained from a medical study by Kandoth et al. (2013) is considered. This table comprises information about $n = 9786$ biological samples and its joint frequencies correspond to the number cross-classified cases of the categorical variables cancer type (U), which has $r = 11$ categories, and significantly mutated gene (V), which has

$c = 127$ categories. The set of genes are divided into 20 groups, which are defined according to biological features. This contingency table can be obtained from the supplementary material in Kandoth et al. (2013) and it is also included in Tables 8 and 9 in the Appendix. In Table 3 we show, for each of the groups $g = 1, \ldots, 20$: its description according to Kandoth et al. (2013), a color to represent it in the upcoming results, the number of genes (categories) in each group ($\mathcal{S}_g$), and the percentage of cells the contingency table within each group which are sparse (that is, for each $i = 1, \ldots, 11$ the cardinality of $\{o_{ij} : o_{ij} < 5,\ j = 1, \ldots, \mathcal{S}_g\}$ divided by $\mathcal{S}_g$ times 100). We point out the noticeable amount of joint frequencies which are below the usual threshold of 5, being the level of sparsity greater than 49% within all the groups. For the interpretation of references to color in Table 3, the reader is referred to the web version of this article.

Table 3 Main features of the contingency table from Kandoth et al. (2013) used to illustrate our approach

Group (g)	Description	Color	Size ($\mathcal{S}_g$)	Sparsity (%)
1	Transcription factor/regulator	🟪	21	75.3
2	Histone modifier	🟩	13	49.0
3	Genome integrity	🟦	13	51.7
4	RTK signalling	🟫	9	58.6
5	Cell cycle	🟥	7	74.0
6	MAPK signalling	⬛	7	64.9
7	PI(3)K signalling	🟧	6	57.6
8	TGF-β signalling	🟪	5	80.0
9	Wnt/β-catenin signalling	⬜	5	70.9
10	Histone	🟩	3	97.0
11	Proteolysis	🟨	3	69.7
12	Splicing	🟥	3	78.8
13	HIPPO signalling	🟦	2	81.8
14	DNA methylation	🟥	2	59.1
15	Metabolism	🟪	2	86.4
16	NFE2L	🟥	2	68.2
17	Protein phosphatase	🟨	2	72.7
18	Ribosome	🟨	2	86.4
19	TOR signalling	🟪	2	63.6
20	Other	🟧	18	64.1

Table 4 Assignment of genes in T_0 to clusters in T solving model (2)–(6) (Part I)

χ^2 ($p-$value)	Cluster 1	Cluster 2	Cluster 3	Cluster 4
3402.28 (< 0.000)	2 3 4 5 6 7 9 10 11 13 14 16 17 18 19 21 22 23 24 27 28 30 31 32 33 34 35 36 37 38 39 40 42 43 45 46 47 48 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 95 96 97 102 103 104 106 107 108 109 110 111 112 113 114 115 116 117 119 120 121 122 123 124 125 126	1 8 12 15 20 25 26 29 41 44 49 68 94 98 99 100 101 105 118 127		
6768.98 (< 0.000)	2 3 4 5 6 7 9 10 11 13 14 15 16 17 18 19 21 22 23 24 27 28 30 31 32 33 34 35 36 37 38 39 40 42 43 44 45 46 47 48 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 95 96 97 102 103 104 106 107 108 109 110 111 112 113 114 115 116 117 119 120 121 122 123 124 125 126	8 12 20 49 68 94 98 100 101 105 118 127	1 25 26 29 41 99	
8532.66 (< 0.000)	3 4 6 7 9 10 13 15 17 18 19 22 23 27 28 30 33 34 35 36 37 38 39 40 42 43 44 45 46 47 48 50 51 52 53 55 56 57 58 59 60 62 63 64 65 67 70 74 75 77 78 79 80 81 82 84 85 87 88 90 91 95 97 102 108 109 110 111 112 113 114 115 117 119 120 121 122 123 124 125	2 5 11 14 16 21 24 31 32 54 61 66 69 71 72 73 76 83 86 89 92 93 96 103 104 106 107 116 126	8 12 20 49 68 94 98 99 100 101 105 118	1 25 26 29 41 127

Table 5 Assignment of genes in T_0 to clusters in T solving model (2)–(6) (Part II)

χ^2 (p−value)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
10278.21 (< 0.000)	3 4 6 7 9 10 13 15 19 22 23 27 28 30 33 34 35 36 37 38 39 40 42 43 44 45 46 47 48 50 51 52 53 55 56 57 58 59 60 62 63 65 67 70 74 75 78 79 81 84 85 87 88 90 91 97 102 103 107 108 109 110 111 112 113 114 115 117 119 120 121 122 123 124 125	2 5 11 14 16 21 24 31 32 54 61 66 69 71 72 73 76 83 86 89 92 93 96 104 106 116 126	8 12 20 49 94 98 99 100 101 105 118	17 18 64 68 77 80 82 95	1 25 26 29 41 127					
11483.64 (< 0.000)	3 4 6 7 9 13 15 22 23 27 28 30 33 34 35 36 37 38 39 40 42 44 45 47 50 51 52 53 55 56 57 58 59 62 63 65 67 70 74 75 78 79 81 84 85 87 88 90 91 97 102 103 108 109 110 111 112 113 114 115 117 119 120 121 122 123 124 125 126	5 11 21 24 31 32 43 48 54 61 72 73 83 86 89 92 104 106 107	2 10 14 16 19 46 60 66 69 71 76 93 96 116	8 12 20 49 94 98 99 100 101 105 118	17 18 64 68 77 80 82 95	1 25 26 29 41 127				
12446.15 (< 0.000)	3 4 6 7 9 13 15 22 23 27 28 30 33 34 36 37 38 39 40 44 45 46 47 48 50 51 52 53 56 57 58 62 65 67 70 74 75 78 79 84 85 88 91 93 97 102 103 107 108 109 110 111 112 113 114 115 117 119 120 121 123 124 125	5 11 21 24 31 32 43 54 61 72 73 83 86 89 90 92 104 106 126	2 8 10 14 16 19 60 66 69 71 76 96 116	12 20 49 94 98 99 100 101 105 118	17 18 64 68 77 80 81 82 95	35 42 55 59 63 87 122	1 25 26 29 41 127			
13064.41 (< 0.000)	3 6 7 9 13 22 28 34 36 37 38 39 40 44 45 46 48 50 51 52 53 58 65 67 70 74 75 79 84 85 88 91 107 108 109 110 112 115 117 119 120 121 124 125	5 11 21 24 31 32 43 54 61 72 73 83 86 89 90 92 104 106 114 126	4 23 27 30 33 47 56 57 62 78 97 102 103 111 113 123	2 10 14 15 16 19 60 66 69 71 76 93 96 116	8 12 20 49 94 98 99 100 101 105 118	17 18 64 68 77 80 81 82 95	35 42 55 59 63 87 122	1 25 26 29 41 127		
13552.2 (< 0.000)	3 7 9 13 22 28 36 37 38 39 40 44 45 46 50 51 52 53 58 65 67 70 74 75 79 85 88 91 108 109 110 112 115 117 119 120 121 124 125	5 6 11 21 24 31 32 43 54 61 83 84 86 89 90 92 104 106 114 126	4 23 27 30 33 47 56 57 62 78 97 102 103 111 113 123	2 10 14 15 16 19 60 66 69 71 76 93 96 116	8 12 20 49 94 98 99 100 101 105 118	17 18 64 68 77 80 81 82 95	35 42 55 59 87 122	1 25 26 29 41 127	34 48 63 72 73 107	
13902.5 (< 0.000)	6 7 9 13 22 31 33 36 38 39 43 44 46 55 70 78 79 81 84 85 90 93 103 108 112 114 115 117 121 122 123 125 126	5 11 21 24 32 34 54 61 72 73 83 86 89 92 104 106	3 28 50 51 53 67 74 75 88 91 109 110 119 124	2 10 14 15 16 19 60 66 69 71 76 96 116	4 23 27 30 47 56 57 62 97 102 111 113	8 12 20 49 94 98 99 100 101 105 118	37 40 45 48 52 58 63 65 107 120	17 18 64 68 77 80 82 95	1 25 26 29 41 127	35 42 59 87

In what follows, we present results obtained from clustering the contingency table of Kandoth et al. (2013) using two different clustering structures. The optimization models involved in the experiments have been solved using Bonmin (Bonami and Lee 2017) under Pyomo. Bonmin is an open-source numerical optimization procedure for solving general Mixed Integer Nonlinear Programs by means of Branch-and-Bound and Branch-and-Cut algorithms, thus avoiding an explicit complete enumeration of all the feasible solutions of the models stated in Sect. 3. In order to avoid being stuck

Table 6 Assignment of genes in T_0 to clusters in T solving model (2)–(7) (Part I)

χ^2 (p−value)	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1185.51 (< 0.000)	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 40 43 44 46 48 49 50 51 52 53 54 55 56 60 61 64 65 66 67 68 69 70 71 72 73 76 77 79 80 81 82 83 84 86 90 92 93 94 95 96 98 99 100 101 105 106 107 118 122 126 127	22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 41 42 45 47 57 58 59 62 63 74 75 78 85 87 88 89 91 97 102 103 104 108 109 110 111 112 113 114 115 116 117 119 120 121 123 124 125		
3230.99 (< 0.000)	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 37 40 41 43 44 46 48 49 50 51 52 53 54 55 56 58 60 61 64 65 66 67 68 69 70 71 72 73 76 83 85 86 92 93 96 104 106 107 108 116 125 126	22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 38 39 42 45 47 57 59 62 74 75 77 78 79 80 81 82 84 87 88 89 90 91 95 97 102 103 109 110 111 112 113 114 115 117 119 120 121 122 123 124	63 94 98 99 100 101 105 118 127	
4380.75 (< 0.000)	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 36 40 43 46 48 49 50 51 52 53 54 55 56 58 60 61 64 65 66 67 68 69 70 71 72 73 76 77 79 80 81 82 83 84 86 90 92 93 95 96 104 106 107 116 122 126	22 23 24 25 26 27 28 29 30 31 32 33 34 37 38 39 41 44 45 47 57 62 74 75 85 88 89 91 97 102 103 108 109 110 111 112 113 114 115 117 119 120 121 123 124 125	63 94 98 99 100 101 105 118 127	35 42 59 78 87

Table 7 Assignment of genes in T_0 to clusters in T solving model (2)–(7) (Part II)

χ^2 ($p-$value)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
5370.6 (< 0.000)	22 23 24 25 26 27 28 29 30 31 32 33 34 36 37 38 39 40 41 44 45 47 57 62 74 75 88 91 97 102 108 109 110 111 112 113 114 115 117 119 120 121 123 124	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 60 71 76 85 93 96 103 107 116 125 126	35 42 46 48 49 50 51 52 53 54 55 56 58 59 64 65 66 67 68 69 70 77 78 79 80 81 82 84 87 90 95 122	43 61 72 73 83 86 89 92 104 106	63 94 98 99 100 101 105 118 127					
6520.47 (< 0.000)	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 36 40 44 46 48 49 50 51 52 53 54 55 56 58 60 64 65 66 67 68 69 70 71 76 79 81 84 85 90 93 96 103 107 116 122 125 126	22 23 24 25 26 27 28 29 30 31 32 33 34 37 38 39 41 45 47 57 62 74 75 88 91 97 102 108 109 110 111 112 113 114 115 117 119 120 121 123 124	43 61 72 73 83 86 89 92 104 106	63 94 98 99 100 101 118 127	77 80 82 95 105	35 42 59 78 87				
7144.85 (< 0.000)	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 36 40 44 48 49 50 51 52 53 54 55 56 58 64 65 66 67 68 69 70 79 81 84 90 93 107 108 122 125 126	22 23 24 25 26 27 28 29 30 31 32 33 34 37 38 39 41 45 47 57 62 74 75 85 88 91 97 102 103 109 110 111 112 113 114 115 117 119 120 121 123 124	43 61 72 73 83 86 89 92 104 106	63 94 98 99 100 101 118 127	46 60 71 76 96 116	77 80 82 95 105	35 42 59 78 87			
7583.56 (< 0.000)	22 23 24 25 26 27 28 29 30 31 32 33 34 38 39 45 47 57 62 74 75 85 89 91 97 102 103 109 110 111 112 113 114 115 117 119 120 121 123 124 125	36 37 40 48 49 50 51 52 53 54 55 56 58 64 65 66 67 68 69 70 79 81 84 88 90 107 122	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 41 108 126	43 61 72 73 83 86 92 104 106	44 98 99 100 101 105 118 127	46 60 71 76 93 96 116	63 77 80 82 94 95	35 42 59 78 87		
8011.38 (< 0.000)	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 41 79 108 126	22 23 24 25 26 27 28 29 30 31 32 33 34 36 38 45 47 62 85 93 103 125	37 40 48 49 50 51 52 53 54 55 56 58 64 65 66 67 68 69 70 105 107 120	39 57 74 75 88 91 97 102 109 110 111 112 113 114 115 117 119 121 123 124	43 61 72 73 83 86 89 92 104 106	46 77 80 81 82 84 90 95 122	63 94 98 99 100 101 118 127	44 60 71 76 96 116	35 42 59 78 87	
8111.06 (< 0.000)	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 41 85 108 126	38 39 45 57 78 89 90 93 97 102 103 104 111 112 113 114 115 117 120 121 122 123	37 48 49 50 51 52 53 54 55 56 58 64 65 66 67 68 69 70 79 81 107	22 23 24 25 26 27 28 29 30 31 32 33 34 36 47 62	74 75 88 91 109 110 119 124 127	44 94 98 99 100 101 105 118	43 61 72 73 83 86 92 106	46 77 80 82 84 95 125	40 60 71 76 96 116	35 42 59 63 87

at local optima, each problem has been solved 10 times with a time limit of one hour in a PC Intel® Core$^{\text{TM}}$ i7-7700, 16GB of RAM.

As abovementioned, the contingency table T_0 in Kandoth et al. (2013) is highly sparse. It is well known that in such case the asymptotic distribution of the χ^2 statistic fails and thus statistical conclusions about the statistical dependence between U and V cannot be inferred (Agresti and Yang 1987). In order to overcome such limitation, a less granular representation of the genes can be considered so that the columns of table T_0 (genes) are clustered into broader categories (groups of genes) in such a way that the aggregated joint frequencies are larger than a threshold, and thus an eventual statistical dependence between U and V in the original table T_0 could be revealed. To do so, the optimization model defined by (2)–(6) is solved for $\beta = 5$. Tables 4 and 5 contain the assignment of the genes in T_0 to the clusters (new categories made up of groups of genes) in T for the number of clusters k varying from 2 to 10. Thus, a initial table T_0 with $r = 11$ and $c = 127$ is reduced to tables T of $r = 11$ and

$c = k$, for $k = 2, \ldots, 10$. First column of Tables 4 and 5 contains the values of the χ^2 statistic in T (and the associated $p-$value in parentheses). We point out that statistical dependence between U and V is detected when the granularity of V is fixed to $k = 2, \ldots, 10$ and a significance level of $\alpha = 5\%$. Thus, we can conclude that, under the aforementioned conditions, there does not exist a extreme grouping since the null hypothesis of statistical independence is rejected for all $k \geq 2$.

The genes in T_0 are split into 20 groups defined through biological features. A plausible requirement could be that the genes of some groups in T_0 must belong to the same clusters in T to avoid having genes of the same groups spread out across different clusters as in Tables 4 and 5. These must-link conditions can be imposed in a preprocessing step. In this case, we impose that genes in Groups 1, 2, 4 and 6, respectively, must belong to the same cluster. In addition, our preprocessing incorporates the requirement that Groups 4 and 6 belong to the same cluster. Conversely, some groups might be required to belong to different clusters. This structure is illustrated by imposing constraint (7) for $j \in$ Group 1 and $j' \in$ Group 2. Tables 6 and 7 contain the assignment of the genes in T_0 to the clusters (new categories) in T for the number of clusters k varying from 2 to 10. As before, the first column of such tables contains the values of the χ^2 statistic in the clustered table T (and the associated $p-$value in parentheses). In this case, statistical dependence between U and V is also detected when the granularity of V is fixed to $k = 2, \ldots, 10$, a significance level of $\alpha = 5\%$ is considered as well as the group structures in the clustering process. Thus, we can conclude that, under the aforementioned conditions, there does not exist a extreme grouping since the null hypothesis of statistical independence is rejected for all $k \geq 2$.

The contingency tables obtained form the clusterings shown in Tables 4, 5, 6 and 7 are depicted in the Supplementary Material. We refer the reader to the web version of this article for the interpretation of references to colors in Tables 4-7.

5 Conclusions

In this paper we have addressed the problem of clustering categories in contingency tables maximizing the χ^2 statistic (Mirkin 2001; Pearson 1900). Solving this clustering problem for different sizes, namely different granularities of the categorical variables under study, allows us to identify extreme groupings or, in other words, the way categories can be clustered into larger ones so that the dependence of the variables is no longer detected if the granularity of the variables is reduced. To do so, a combinatorial mathematical optimization model has been stated, which allows to accommodate structural properties of the clusters in the clustered table which naturally arise in the context of the dataset under study. An assignment formulation has been proposed for such model, namely (CCT), yielding a 0–1 nonlinear optimization problem. Requirements on the clusters, such as non-sparsity conditions, relational and cannot link constraints, have been stated as linear constraints. In addition, a set partitioning reformulation of (CCT) is also proposed. Our methodology is illustrated using a dataset in a medical study, which naturally demands the use of the tools proposed in this paper to handle the study of statistical dependence between its variables under structural conditions on the clusters.

The problem studied in this paper can be extended in a few directions. First, other criteria different from χ^2 could be considered to measure statistical dependence (Goodman and Kruskal 1979; Joe 1989). Second, exploring different criteria to group the categories in contingency tables different from statistical dependence, and which are defined through appropriate combinatorial optimization models, could be also explored as an extension to this paper. Some interesting examples could be to explore patterns in the observed joint frequencies to group the categories in a contingency table, or to identify those patterns in the coordinates given by Correspondence Analysis (Ciampi et al. 2005; Pledger and Arnold 2014; van de Velden et al. 2020). Third, tighter formulations for (CCT) could be explored, in combination with metaheuristic approaches such as the Variable Neighborhood Search (Mladenović and Hansen 1997) or the Large Neighborhood Search (Pisinger and Ropke 2010), to address larger tables. Finally, extensions of the proposed methodology for dealing with multi-way tables require further research (Agresti and Gottard 2007).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-022-00508-4.

Acknowledgements The authors would like to thank the reviewers for their comments on prior versions of this manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214), FQM-329, P18-FR-2369 and US-1381178 (Junta de Andalucía, with FEDER Funds), PID2019-110886RB-I00 and PID2019-104901RB-I00 (funded by MCIN/AEI/10.13039/501100011033). This support is gratefully acknowledged.

Appendix

Tables 8 and 9 contain the transposed contingency table T_0 used in the illustrations of our methodology in Sect. 4 (genes in the rows and cancer types in the columns). This table is obtained from the supplementary material in Kandoth et al. (2013). First column identifies the groups of genes, which are also colored in different ways, whereas second and third contain a numerical and alphanumerical identifier for the genes, respectively. The remaining 11 columns include the joint frequencies of the variables genes and cancer type.

Table 8 Transposed contingency table T_0 used in Sect. 4 (Part I) (Kandoth et al. 2013)

			BLCA	BRCA	COAD/READ	GBM	HNSC	KIRC	AML	LUAD	LUSC	OV	UCEC
Group 1	1	VHL	0	0	0	0	0	218	0	0	1	0	2
	2	GATA3	8	81	2	0	6	0	0	6	5	1	1
	3	TSHZ3	2	5	6	2	4	5	1	34	11	3	9
	4	EP300	17	6	4	1	24	6	0	2	8	1	12
	5	CTCF	2	18	3	0	10	2	1	3	0	1	38
	6	TAF1	2	8	3	4	7	5	0	9	12	5	20
	7	TSHZ2	4	7	6	7	4	3	0	15	6	3	4
	8	RUNX1	1	25	2	0	2	0	18	1	0	0	3
	9	MECOM	5	4	2	4	5	4	0	8	8	2	7
	10	TBX3	3	18	2	0	2	0	0	10	5	3	3
	11	SIN3A	1	4	1	2	2	2	0	4	5	2	12
	12	WT1	0	1	2	2	0	3	12	8	4	0	1
	13	EIF4A2	2	4	5	0	0	3	0	4	2	2	3
	14	FOXA1	4	13	0	3	2	0	0	1	1	0	0
	15	PHF6	3	3	0	1	1	2	6	2	2	1	3
	16	CBFB	1	16	0	0	0	1	2	1	1	0	1
	17	SOX9	0	1	8	3	2	3	0	3	1	0	1
	18	ELF3	8	1	7	0	1	0	0	1	0	1	1
	19	VEZF1	2	7	0	2	2	0	0	2	3	0	0
	20	CEBPA	0	0	0	0	0	1	13	0	1	0	0
	21	FOXA2	1	0	1	0	0	0	0	0	0	2	11
Group 2	22	MLL3	24	49	5	9	22	15	1	42	27	6	12
	23	MLL2	25	12	3	5	54	13	1	20	35	2	19
	24	ARID1A	27	15	11	2	9	12	1	14	11	3	69
	25	PBRM1	6	3	0	2	7	137	0	4	6	1	6
	26	SETD2	6	9	5	5	7	48	1	18	5	6	6
	27	NSD1	6	2	1	1	32	4	0	7	9	2	13
	28	SETBP1	2	3	3	4	9	6	2	29	9	0	5
	29	KDM5C	1	4	1	2	3	27	0	11	5	6	5
	30	KDM6A	26	8	0	3	8	4	3	2	7	0	2
	31	MLL4	7	5	4	6	8	4	0	4	7	1	19
	32	ARID5B	3	3	0	1	10	3	0	5	3	2	22
	33	ASXL1	3	3	3	0	9	4	5	3	9	0	2
	34	EZH2	1	1	0	3	1	3	3	5	4	0	4
Group 3	35	TP53	49	251	113	82	210	9	15	118	138	299	64
	36	ATM	11	16	11	4	8	12	0	18	7	4	15
	37	ATRX	8	9	2	16	13	8	0	14	10	2	7
	38	BRCA2	6	13	3	4	11	8	0	13	10	10	10
	39	ATR	4	6	4	4	16	5	0	13	7	2	16
	40	STAG2	10	7	2	12	2	7	6	6	6	3	9
	41	BAP1	4	2	0	2	3	42	0	3	1	2	5
	42	BRCA1	4	12	0	3	8	4	0	8	9	11	2
	43	SMC1A	3	6	3	5	3	2	7	3	1	4	10
	44	SMC3	1	3	0	4	5	5	7	6	4	1	1
	45	CHEK2	2	3	0	5	7	3	0	2	2	1	3
	46	RAD21	2	4	2	1	3	0	5	6	2	1	2
	47	ERCC2	12	1	1	0	1	1	0	3	0	1	1
Group 4	48	EGFR	1	5	3	77	14	7	2	26	5	6	3
	49	FLT3	2	3	0	5	2	2	53	9	7	3	2
	50	EPHA3	1	4	6	3	11	2	1	20	11	3	5
	51	ERBB4	2	6	7	1	13	6	0	17	9	0	6
	52	PDGFRA	6	3	2	11	3	6	1	15	7	3	3
	53	EPHB6	3	3	0	4	4	5	0	22	6	1	4
	54	FGFR2	2	7	0	1	2	1	0	7	4	0	24
	55	KIT	1	4	2	3	3	3	8	4	6	6	5
	56	FGFR3	8	1	1	4	5	6	0	1	4	1	1
Group 5	57	CDKN2A	4	0	1	2	64	4	0	15	26	0	1
	58	RB1	14	14	1	24	9	1	0	12	12	6	9
	59	CDK12	4	7	3	1	5	6	0	7	1	9	5
	60	CDKN1B	2	7	2	1	2	0	0	4	0	1	2
	61	CCND1	2	1	0	0	1	0	0	2	1	0	12
	62	CDKN1A	12	0	0	1	0	1	0	1	2	1	0
	63	CDKN2C	0	0	0	3	0	1	0	0	1	2	0

Table 9 Transposed contingency table T_0 used in Sect. 4 (Part II) (Kandoth et al. 2013)

			BLCA	BRCA	COAD/READ	GBM	HNSC	KIRC	AML	LUAD	LUSC	OV	UCEC
Group 6	64	KRAS	0	6	87	2	1	1	8	60	2	2	46
	65	NF1	7	19	2	32	8	7	2	27	18	12	8
	66	MAP3K1	3	55	0	6	3	5	0	4	3	1	8
	67	BRAF	2	3	7	6	3	1	0	15	8	2	2
	68	NRAS	2	1	17	1	0	0	15	4	1	2	6
	69	MAP2K4	0	31	5	0	1	0	0	3	1	1	3
	70	MAPK8IP1	2	2	4	2	2	2	0	4	2	1	1
Group 7	71	PIK3CA	17	256	34	32	62	12	0	10	26	2	120
	72	PTEN	3	29	2	89	4	18	0	5	14	2	146
	73	PIK3R1	1	19	4	33	5	2	0	3	1	1	71
	74	TLR4	2	9	0	1	6	2	1	26	10	3	1
	75	PIK3CG	2	3	1	7	8	3	0	12	13	3	3
	76	AKT1	0	19	0	1	2	2	0	0	1	0	3
Group 8	77	SMAD4	2	3	19	1	6	2	0	7	5	0	0
	78	TGFBR2	3	3	5	2	9	1	0	2	3	3	3
	79	ACVR1B	0	5	7	0	4	4	0	5	2	1	4
	80	SMAD2	1	4	11	0	3	2	0	2	2	0	3
	81	ACVR2A	1	4	5	0	2	1	0	2	2	0	1
Group 9	82	APC	4	4	158	1	12	6	0	21	7	7	13
	83	CTNNB1	2	1	9	1	2	1	0	8	3	2	65
	84	AXIN2	3	1	7	1	5	1	0	2	1	1	6
	85	TBL1XR1	2	8	0	0	3	3	0	5	2	1	3
	86	SOX17	0	0	1	1	1	0	0	1	0	0	7
Group 10	87	HIST1H1C	1	3	2	2	4	1	0	1	1	4	0
	88	H3F3C	0	0	0	2	2	0	0	4	2	0	2
	89	HIST1H2BD	1	0	0	0	4	0	0	0	2	1	6
Group 11	90	FBXW7	9	6	22	1	15	1	0	3	9	3	27
	91	KEAP1	3	1	0	0	12	2	0	39	21	1	3
	92	SPOP	1	1	0	0	3	0	0	1	1	1	15
Group 12	93	SF3B1	9	14	2	2	2	4	1	5	4	0	5
	94	U2AF1	1	2	1	0	4	0	8	6	0	0	2
	95	PCBP1	1	0	5	0	0	1	0	1	0	0	2
Group 13	96	CDH1	5	53	1	1	4	2	0	3	3	1	7
	97	AJUBA	2	1	0	1	18	2	0	2	0	0	0
Group 14	98	DNMT3A	0	4	2	0	5	5	51	9	7	3	3
	99	TET2	3	3	0	2	1	8	17	7	4	0	5
Group 15	100	IDH1	3	2	0	15	1	2	19	2	2	0	2
	101	IDH2	0	0	3	0	0	0	20	1	0	0	1
Group 16	102	NFE2L2	9	1	0	0	16	5	0	5	26	0	12
	103	NFE2L3	3	6	0	1	4	1	0	0	4	1	4
Group 17	104	PPP2R1A	1	1	3	0	4	5	0	3	8	4	20
	105	PTPN11	0	1	2	5	1	1	9	6	3	1	2
Group 18	106	RPL22	0	0	0	1	2	2	0	1	0	0	25
	107	RPL5	0	3	0	8	0	6	0	1	2	0	2
Group 19	108	MTOR	2	11	7	4	4	25	0	17	8	6	12
	109	STK11	0	2	0	0	1	1	0	20	3	0	1
Group 20	110	NAV3	5	11	4	3	22	6	0	49	33	4	12
	111	NOTCH1	5	3	0	0	58	4	1	7	14	2	4
	112	LRRK2	5	5	5	3	15	6	0	15	20	9	8
	113	MALAT1	15	8	0	0	19	8	0	22	10	3	0
	114	ARHGAP35	5	7	1	2	11	5	1	9	10	5	23
	115	POLQ	7	6	1	3	13	5	0	13	16	3	9
	116	NCOR1	8	30	1	2	10	3	0	6	6	1	3
	117	USP9X	3	9	0	2	13	4	1	12	8	1	15
	118	NPM1	0	0	0	1	1	0	54	2	0	0	1
	119	HGF	1	4	0	1	8	1	0	24	10	2	3
	120	EPPK1	2	2	0	8	8	3	0	7	7	1	7
	121	AR	1	5	4	0	7	2	0	4	6	1	8
	122	LIFR	1	6	10	0	8	2	0	2	3	2	4
	123	PRX	5	4	1	2	5	5	1	1	2	1	3
	124	CRIPAK	2	2	0	1	3	2	0	12	0	0	1
	125	EGR3	1	2	0	1	0	1	0	1	2	0	1
	126	B4GALT3	0	1	1	0	0	1	0	0	2	0	2
	127	MIR142	0	1	0	0	0	0	4	0	0	0	0

References

Abin AA (2019) Clustering in the presence of side information: a non-linear approach. Int J Intel Comput Cybern 12(2):292–314

Agresti A, Gottard A (2007) Independence in multi-way contingency tables: S.N. Roy's breakthroughs and later developments. J Stat Plan Inference 137(11):3216–3226

Agresti A, Yang MC (1987) An empirical investigation of some effects of sparseness in contingency tables. Comput Stat Dat Anal 5:9–21

Ailem M, Role F, Nadif M (2016) Graph modularity maximization as an effective method for co-clustering text data. Knowl-Based Syst 109:160–173

Ailem M, Role F, Nadif M (2017) Model-based co-clustering for the effective handling of sparse data. Pattern Recogn 72:108–122

Ailem M, Role F, Nadif M (2017) Sparse Poisson latent block model for document clustering. IEEE Trans Knowl Data Eng 29(7):1563–1576

Álvarez de Toledo P, Núñez F, Usabiaga C (2018) Matching and clustering in square contingency tables. Who matches with whom in the Spanish labour market. Comput Stat Dat Anal 127:135–159

Anderson R, Huchette J, Ma W, Tjandraatmadja C, Vielma JP (2020) Strong mixed-integer programming formulations for trained neural networks. Math Program 183:3–39

Baesens B, Setiono R, Mues C, Vanthienen J (2003) Using neural network rule extraction and decision tables for credit-risk evaluation. Manage Sci 49(3):312–329

Benati S, García S (2014) A mixed integer linear model for clustering with variable selection. Comput Oper Res 43:280–285

Bertsimas D, King A (2016) OR forum - An algorithmic approach to linear regression. Oper Res 64(1):2–16

Bertsimas D, Shioda R (2007) Classification and regression via integer optimization. Oper Res 55(2):252–271

Blanquero R, Carrizosa E, Molero-Río C, Romero Morales D (2020) Sparsity in optimal randomized classification trees. Eur J Oper Res 284(1):255–272

Blyth CR (1972) On simpson's paradox and the sure-thing principle. J Am Stat Assoc 67(338):364–366

Bock HH (2003) Two-way clustering for contingency tables: maximizing a dependence measure. In: Between data science and applied data analysis, Springer, Heidelberg, Germany, pp 143–154

Bonami P, Lee J (June 2017) Bonmin user's manual. Technical report, IBM Corporation

Bottou L, Curtis F, Nocedal J (2018) Optimization methods for large-scale machine learning. SIAM Rev 60(2):223–311

Boulle M (2004) Khiops: A statistical discretization method of continuous attributes. Mach Learn 55(1):53–69

Carrizosa E, Guerrero V (2014) rs-Sparse principal component analysis: A mixed integer nonlinear programming approach with VNS. Comput Oper Res 52:349–354

Carrizosa E, Romero Morales D (2013) Supervised classification and mathematical optimization. Comput Oper Res 40(1):150–165

Carrizosa E, Mladenović N, Todosijević R (2013) Variable neighborhood search for minimum sum-of-squares clustering on networks. Eur J Oper Res 230(2):356–363

Carrizosa E, Guerrero V, Romero Morales D (2017a) Visualizing proportions and dissimilarities by space-filling maps: a large neighborhood search approach. Comput Oper Res 78:369–380

Carrizosa E, Nogales-Gómez A, Romero Morales D (2017b) Clustering categories in support vector machines. Omega 66:28–37

Carrizosa E, Guerrero V, Romero Morales D (2018a) On mathematical optimization for the visualization of frequencies and adjacencies as rectangular maps. Eur J Oper Res 265(1):290–302

Carrizosa E, Guerrero V, Romero Morales D (2018b) Visualizing data as objects by DC (difference of convex) optimization. Math Program 169:119–140

Carrizosa E, Guerrero V, Romero Morales D (2019) Visualization of complex dynamic datasets by means of mathematical optimization. Omega 86:125–136

Carrizosa E, Romero Morales V, Guerrero D, Satorra A (2020) Enhancing interpretability in factor analysis by means of mathematical optimization. Multivar Behav Res 55(5):748–762

Carrizosa E, Molero-Río C, Romero Morales D (2021) Mathematical optimization in classification and regression trees. TOP 29(1):5–33

Carrizosa E, Kurishchenko K, Marín A, Romero Morales D (2022) Interpreting clusters via prototype optimization. Omega 107(102543):1–13

Ciampi A, González Marcos A, Castejón Limas M (2005) Correspondence analysis and two-way clustering. SORT 29(1):27–42

Cunningham JP, Ghahramani Z (2015) Linear dimensionality reduction: Survey, insights, and generalizations. J Mach Learn Res 16:2859–2900

Fischetti M, Jo J (2018) Deep neural networks and mixed integer linear optimization. Constraints 23:296–309

Fossier S, Riverain P, Nadif M (2022) Semi-supervised latent block model with pairwise constraints. Mach Learn 111(5):1739–1764

Freling R, Romeijn HE, Romero Morales D, Wagelmans APM (2003) A branch-and-price algorithm for the multiperiod single-sourcing problem. Oper Res 51(6):922–939

Gambella C, Ghaddar B, Naoum-Sawaya J (2021) Optimization problems for machine learning: A survey. Eur J Oper Res 290(3):807–828

Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision-making and a "right to explanation". AI Mag 38(3):50–57

Goodman LA, Kruskal WH (1979) Measures Of Association For Cross Classifications. Springer, New York

Govaert G (1995) Simultaneous clustering of rows and columns. Control Cybern 24(4):437–458

Govaert G, Nadif M (2007) Clustering of contingency table and mixture model. Eur J Oper Res 183(3):1055–1066

Govaert G, Nadif M (2010) Latent block model for contingency table. Comnun Stat Theor Meth 39(3):416–425

Govaert G, Nadif M (2018) Mutual information, phi-squared and model-based co-clustering for contingency tables. Adv Data Anal Classif 12:455–488

Greenacre MJ (1988) Clustering the rows and columns of a contingency table. J Classif 5:39–51

Hansen P, Jaumard B (1997) Cluster analysis and mathematical programming. Math Program 79:191–215

Hochbaum DS, Liu S (2018) Adjacency-clustering and its application for yield prediction in integrated circuit manufacturing. Oper Res 66(6):1571–1585

Joe H (1989) Relative entropy measures of multivariate dependence. J Am Stat Assoc 84(405):157–164

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L (2013) Mutational landscape and significance across 12 major cancer types. Nature 502(7471):333–352

Kerber R (1992) Chimerge: Discretization of numeric attributes. In: Proceedings of the 10th National Conference on Artificial intelligence, pp 123–128

Labiod L, Nadif M (2011) Co-clustering for binary and categorical data with maximum modularity. In: IEEE 11th International conference on Data Mining, IEEE, pp 1140–1145

Mirkin B (2001) Eleven ways to look at the chi-squared coefficient for contingency tables. Am Stat 55(2):111–120

Mladenović N, Hansen P (1997) Variable neighborhood search. Comput Oper Res 24(11):1097–1100

Olafsson S, Li X, Wu S (2008) Operations research and data mining. Eur J Oper Res 187(3):1429–1448

Park K, Lee K, Park S, Lee H (2000) Telecommunication node clustering with node compatibility and network survivability requirements. Manage Sci 46(3):363–374

Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50(302):157–175

Pisinger D, Ropke S (2010) Large neighborhood search. In: Gendreau M, Potvin JY (eds) Handbook of metaheuristics, vol 146, chapter 13, Springer, US, pp 399–419

Pledger S, Arnold R (2014) Multivariate methods using mixtures: correspondence analysis, scaling and pattern-detection. Comput Stat Data Anal 71:241–261

Sağlam B, Salman FS, Sayın S, Türkay M (2006) A mixed-integer programming approach to the clustering problem with an application in customer segmentation. Eur J Oper Res 173(3):866–879

Shmueli G, Yahav I (2017) The forest or the trees? Tackling Simpson's paradox with classification trees. Prod Oper Manag 27(4):696–716

Śmieja M, Wiercioch M (2017) Constrained clustering with a complex cluster structure. Adv Data Anal Classif 11(3):493–518

Toriello A, Vielma JP (2012) Fitting piecewise linear continuous functions. Eur J Oper Res 219(1):86–95

Tsumoto S (2009) Contingency matrix theory: statistical dependence in a contingency table. Inf Sci 179(11):1615–1627

Ustun B, Rudin C (2016) Supersparse linear integer models for optimized medical scoring systems. Mach Learn 102(3):349–391

van de Velden M, van den Heuvel W, Galy H, Groenen PJF (2020) Retrieving a contingency table from a correspondence analysis solution. Eur J Oper Res 283:541–548