applied sciences

MDPI

Article

Dichotomization of Multilevel Variables to Detect Hidden Associations

Asdrúbal López-Chau [1,*], Lisbeth Rodriguez-Mazahua [2], Farid García-Lamont [3], Maricela Quintana-López [4] and Carlos A. Rojas-Hernández [1]

1 Laboratory of Applied Computing Technologies, Universidad Autónoma del Estado de México, CU UAEM Zumpango, K.M. 3.5, Camino Viejo a Jilotzingo, Valle Hermoso, Zumpango 55600, Estado de México, Mexico
2 Tecnológico Nacional de México, I. T. Orizaba, Av. Oriente 9 852, Col. Emiliano Zapata, Orizaba 94320, Veracruz, Mexico
3 CU UAEM Texcoco, Universidad Autónoma del Estado de México, Jardin Zumpango s/n, Fraccionamiento El Tejocote, Texcoco 56259, Estado de México, Mexico
4 CU UAEM Valle de México, Universidad Autónoma del Estado de México, Boulevard Universitario s/n Valle Escondido, Río San Javier, Cd López Mateos 54500, Estado de México, Mexico
* Correspondence: alchau@uaemex.mx

Abstract: A test of independence is commonly used to determine differences (or associations) between samples in a nominal level measurement. Fisher's exact test and Chi-square test are two of the most widely applied tests of independence used in the data analyses in different areas such as information technologies, biostatistics, psychology and health sciences. In some cases, contingency tables with null entries (also called random zeros) arise, particularly if the number of samples is small, and the variables analyzed are multilevel. This situation becomes a problem because if one or more entries in a contingency table are zero or have small values, then the tests of independence produce unreliable results. In this paper, we propose a method to address that issue. The method merges one or more levels of the variables analyzed to create contingency tables with only one degree of freedom, avoiding applying a test of independence on contingency tables with random zeros. The source code (Python) of the method is publicly available for use. The results obtained using our method give a complete panorama of the associations between the variables of a data set. To show the effectiveness of our approach to find dependencies between variables, we use four data sets publicly available on the Internet.

Keywords: test of independence; Chi-square test; Fisher's exact test; multilevel variables

Citation: López-Chau, A.; Rodriguez-Mazahua, L.; García-Lamont, F.; Quintana-López, M.; Rojas-Hernández, C.A. Dichotomization of Multilevel Variables to Detect Hidden Associations. *Appl. Sci.* **2022**, *12*, 12929. https://doi.org/10.3390/app122412929

Academic Editor: Paolino Di Felice

Received: 11 November 2022
Accepted: 13 December 2022
Published: 16 December 2022

1. Introduction

The test of independence is one of the most common tools of inferential statistics applied to areas such as health, psychology and biostatistics. This type of test is used in many studies to determine differences between samples in a nominal level measurement [1].

For example, in [2] the authors applied the Chi-square test on collected data from three psychiatric hospitals in Ogun state, Nigeria, to show the associations between socio-demographic variables and perceived hindrance of mental services. In [3], the Chi-square test was used to know the dependency between the physical activity and school adaptation in two groups of students. In [4], a test of independence was used to support the evidence of the association between the effect of early treatment responses of subjects with major depressive disorder treated with medicines or placebo. Chen, Haung and Ng [5] applied the Chi-square test to evaluate the association between single-nucleotide polymorphisms (SNPs) with multiple diseases. Adamu et al. [6] used the Chi-square test to determine the dependency of the variables of a data set of various cancer types recorded in northeastern states of Nigeria. In [7], the psychological profile of patients with reactions to placebo during a drug provocation test was evaluated by Fisher's exact test.

On the other hand, for more than three decades the validity of independence tests with contingency tables with small expected cell counts has been extensively researched and discussed [8,9]. In practice, when the number of levels of the variables analyzed is high, the contingency tables with random zeros usually arise, especially if the number of samples is small. In extreme cases, the result of tests can be unreliable, even using Fisher's exact test. Some tools such as Minitab [10] do not print any if the expected cell counts is less than two or three, depending on the number of levels in the variables.

This paper proposes the use of a method that allows the building of contingency tables with one degree of freedom. For this purpose, new synthetic (but human-understandable) categories are created, allowing a direct application of Chi-square or Fisher's exact tests of independence.

Our method executes an exhaustive exploration of the combinations between the different values of variables. This can help to discover hidden associations between variables. The results are shown graphically to assist in the finding's identification process. An implementation of the method is publicly available at http://u.pc.cd/JjCrtalK (accessed on 1 December 2022).

In [11], Sharpe suggests, if possible, avoiding contingency tables of more than one degree of freedom. If the above is not possible, he mentions that one of the methods to identify associations in the data is to use partitioning. The method proposed in this article does an exhaustive search on all possible partitions.

Recently, Zeng et al. [12] developed iSuc-ChiDT, a method to identify succinylation sites in proteins. That method encodes the Chi-square statistical difference table to identify positional attributes in the protein. The iSuc-ChiDT resembles our method; however, the purpose of our method is completely different.

The rest of this paper is organized in four Sections. Section 2 explains the creation of contingency tables and the tests of independence, Chi-square and Fisher's exact test. The proposed method, and the data sets used to test the proposal are shown in Section 3. Section 4 shows the results of the application of our methods on three data sets. Finally, conclusions are presented in the last section.

2. Preliminaries

2.1. Contingency Tables

In simple words, a contingency table is a two-way layout of the number of items in the sample set that fall into different categories [13,14]. It is a cross-classification table in which the association between the variables (or if they are not related to each other) can be seen.

A two-way contingency table is created from a sample by using two qualitative variables (henceforth, we will refer them as V_1 and V_2, respectively). The number of columns corresponds to the levels (distinct values) of V_1 (m), and the number of rows corresponds the levels of V_2 (n). Therefore, the number of entries or cells is $m \times n$.

Table 1 shows an example of a two-way contingency table for two dichotomous variables V_1 and V_2. The levels of V_1 are C_1, C_2, whereas for V_2 they are B_1, B_2. The content of each cell is computed from the sample as follows:

a: The number of elements that simultaneously belong to categories B_1 and C_2.
b: The number of elements that simultaneously belong to categories B_2 and C_1.
c: The number of elements that simultaneously belong to categories B_1 and C_2.
d: The number of elements that simultaneously belong to categories B_2 and C_2.

Table 1. Example of a 2 × 2 contingency table.

V_1	V_2		Total
	C_1	C_2	
B_1	a	c	$a + c$
B_2	b	d	$b + d$
Total	$a + b$	$c + d$	N

The letters a, b, c and d are the observed frequencies. The sum of rows and columns of a contingency table is called the marginal totals. The values $a + c$ and $b + d$ form the marginal rows. The values $a + b$ and $c + d$ are the marginal columns. N is the size of the sample. The degrees of freedom are computed with $df = (m - 1) \times (n - 1)$; therefore, for the example of Table 1, $df = 1$.

Once a two-way contingency table is created, a test of independence can be applied to discover an association between the variables V_1 and V_2. In the next Subsections, we will explain two of the most common tests of independence.

2.2. Chi-Square Test

This test of independence is used to determine whether there is a significant association between two variables [15–18].

The null hypothesis (H_0) of the Chi-square test of independence is that there is no association between the variables. The alternative hypothesis (H_1) assumes there is some association between V_1 and V_2.

The Chi-square statistic is defined as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where

O_i is the number of samples in category (cell) i,

E_i is the expected value of category i, i.e., $E_{i,j} = \frac{T_i T_j}{N}$; T_i is the total of row i, and j is the total of column j.

The value of χ^2 and the degrees of freedom are used to search the p-value in the table of distribution of probability of Chi-square. Typically, if the p-value is greater than 0.05 (level of significance), then V_1 and V_2 have no association (H_0 is accepted); otherwise; the variables have an association (H_0 is rejected).

It is common in practice to consider that the test is reliable if $E_i > 5$ for at least 80% of the number of cells.

2.3. Fisher's Exact Test

In some cases, the expected frequencies are lower than five ($E_i < 5$) for more than 20% of cells in contingency tables. To asses the association between variables in these scenarios, it is convenient to apply Fisher's exact test of independence [19,20].

For this test, the null hypothesis (H_0) is that the relative proportions of V_1 are independent of V_2, i.e., the variables have no association. The alternative hypothesis (H_1) is that V_1 and V_2 have an association.

For Table 1, the probability of H_0 to be true is computed with:

$$p = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{N}{a+b}}$$

where $\binom{n}{k}$ is the number of subsets of k elements chosen from a set with n elements.

Similar to the Chi-square test, if $p < 0.05$ (or other level of significance), then H_1 is accepted, and H_0 is rejected. Otherwise, H_0 is accepted and H_1 rejected.

2.4. Problem Formulation

Consider two categorical variables V_1 and V_2, suppose the following:

- V_1 has L_1 levels;
- V_2 has L_2 levels.

A test of independence between these variables requires a contingency table $\mathbb{T}$ with $(L_1 - 1) \times (L_2 - 1)$ degrees of freedom. The number of entries in $\mathbb{T}$ is $L_1 \times L_2$. A widely accepted rule is that all of the observed and expected frequencies in contingency tables should be at least five; this criteria is used in the Scipy library. Some other tools, such as Minitab, do not display the *p*-value when any expected count is less than one because the results can be invalid.

In some cases, the data analyzed produces contingency tables that contain one or more cells with low values, or even with zero entries. In these scenarios, the results of the Chi -square test and Fisher's exact test are unreliable. To face this problem, in the next section, we propose a simple method that merges levels of variables and apply a brute force approach to explore all the possible combinations. Due to this merging of levels, the contingency tables are less likely to have cells with zero values. This can help in find hidden associations between pairs of variables.

3. Materials and Methods

3.1. Method

To avoid the scenarios, such as the one explained in the problem formulation subsection, we propose to transform the original variables V_1 and V_2, into new synthetic dichotomous variables $\mathcal{A}$ and $\mathcal{B}$, respectively. These new variables contain the fusion of different levels of each variable.

In general, for each synthetic variable ($\mathcal{A}$ or $\mathcal{B}$), the exact number of fusions of levels is the semi-sum of the number of combinations of L_1 or L_2 taking r levels at a time. This is computed as follows:

$$N_{\mathcal{A}} = \left\lfloor \frac{1}{2} \sum_{r=1}^{L_1} \frac{L_1!}{r!(L_1 - r)!} \right\rfloor \tag{1}$$

$$N_{\mathcal{B}} = \left\lfloor \frac{1}{2} \sum_{r=1}^{L_2} \frac{L_2!}{r!(L_2 - r)!} \right\rfloor \tag{2}$$

where

$N_{\mathcal{A}}$ is the number of fusions of the L_1 levels of variable V_1,
$N_{\mathcal{B}}$ is the number of fusions of the L_2 levels of variable V_2, and
$\lfloor x \rfloor$ is the floor of the value x.

Based on these synthetic and dichotomous variables, $\mathcal{A}$ and $\mathcal{B}$, the number of contingency tables that can be created is computed with $N_{\mathcal{B}} \times N_{\mathcal{B}}$. The general structure of these tables is shown in Table 2.

Table 2. Structure of tables of contingency with synthetic variables.

	$\mathcal{B}$		
$\mathcal{A}$	**Category I $_{\mathcal{B}}$**	**Category II $_{\mathcal{B}}$**	**Total**
Category I $_{\mathcal{A}}$	a	c	a + c
Category II $_{\mathcal{A}}$	b	d	b + d
Total	a + b	c + d	N

The Categories $I_{\mathcal{A}}$ and $II_{\mathcal{A}}$ of Table 2 contain one level of V_1 or the merge of more than one level of V_1. These mixtures of levels are mutually exclusive.

For the Categories I_B and II_B is the same but considering the variable V_2. The entries a, b, c and d of Table 2 are computed as was explained for Table 1.

Example 1. *Suppose we have the following two multilevel variables.*

- *V_1 with levels: {single, married, divorced, widower};*
- *V_2 with levels: {Private, self-employment, never-worked}.*

For V_1, we have $L_1 = 4$, whereas for V_2 the number of levels is $L_2 = 3$. The number of possible combinations for the fusions of variables $\mathcal{A}$ and $\mathcal{B}$ is computed with Equations (1) and (2):

$$N_{\mathcal{A}} = \left\lfloor \frac{1}{2} \sum_{r=1}^{L_1} \frac{L_1}{r(L_1 - r)} \right\rfloor = \left\lfloor \frac{1}{2} \left(\frac{4!}{1!(4-1)!} + \frac{4!}{2!(4-2)!} + \frac{4!}{3!(4-3)!} + \frac{4!}{4!(4-4)!} \right) \right\rfloor = 7$$

$$N_{\mathcal{B}} = \left\lfloor \frac{1}{2} \sum_{r=1}^{L_2} \frac{L_2}{r(L_2 - r)} \right\rfloor = \left\lfloor \frac{1}{2} \left(\frac{3!}{1!(3-1)!} + \frac{3!}{2!(3-2)!} + \frac{3!}{3!(3-3)!} \right\rfloor = 3$$

A total of 21 tests of independence are needed to completely explore the possible associations between V_1 and V_2. For this small example, it is possible to show all the fusions of levels; these are listed in Table 3.

Table 3. List of all possible fusions of variables' levels.

$\mathcal{A}$		$\mathcal{B}$	
Category I$_{\mathcal{A}}$	**Category II$_{\mathcal{A}}$**	**Category I$_{\mathcal{B}}$**	**Category II$_{\mathcal{B}}$**
1 single	1 married, divorced, widower	1 private	1 self-employment, never-worked
2 married	2 single, divorced, widower	2 self-employment	2 private, never-worked
3 divorced	3 single, married, widower	3 never-worked	3 private, self-employment
4 widower	4 single, married, widower		
5 single, married	5 divorced, widower		
6 single, divorced	6 married, divorced		
7 single, widower	7 married, divorced		

Algorithm 1 shows the complete method for completely exploring the associations between categorical variables with two or more levels. An implementation of this method in the Python programming language is publicly available at http://u.pc.cd/JjCrtalK (accessed on 1 December 2022).

The computational load of Algorithm 1 is high. Considering only the most costly steps, we have the following. The generation of all combinations of n elements, taking r at a time (Line 26) has a cost of $\mathcal{O}\left(\binom{n}{1} + \binom{n}{2} + ... + \binom{n}{\lfloor \frac{n}{2} \rfloor} \right) \approx \mathcal{O}(2^n)$. In addition, the creation of the synthetic variables, that is, the level partitioning of each variable, has a linear cost with respect to the size of the data set. The Processing with categorical variables of more than 12 levels with hundreds of pieces of data takes several minutes on everyday personal computers and hours if the number of samples are thousands.

Algorithm 1: Exhaustive test of independence for categorical variables

Data: V_1, V_2: Values of categorical variables to test;
α : Level of significance
Result: Tests of independence

```
L1 ← number of levels of V1;
L2 ← number of levels of V2;
A = GenerateCombinations(Levels of V1);
B = GenerateCombinations(Levels of V2);
for each ac ∈ CategoryI_A do
    bd ← Category II_A;
    for each ab ∈ CategoryII_B do
        cd ← Category II_B;
        a ← ac∩ ab;
        b ← bd∩ ab;
        c ← ac∩ cd;
        d ← bd∩ cd;
        Create contingency table with entries a,b,c,d (see Table 2);
        Compute test of independence;
        if p-value< α then a
            Save table of contingency and p-value ;
        end
    end
    return tables of contingency and p-values;
end
GenerateCombinations (Levels):
    n ← number of levels;
    Category I← ∅ //Begins empty;
    Category = II← ∅ //Begins empty;
    for r = 1 to n do
        C = Generate all combinations of n elements, taking r at a time;
        for each combination c_i in C do
            Category I← Category I ⋃ c_i ;
            Category II← Levels - c_i;
        end
    end
    Get rid of repeated (upper half) elements for each category;
    return Category I, Category II
return
```

3.2. Data Sets

Table 4 shows the main characteristics of the data sets used for the experiments; a brief description of these is given below:

- **Scholarship** data set. This is a synthetic data set. It was created by the authors of this article to demonstrate the presented method. It contains four attributes: (a) score, numerical type with values between 5 and 10; (b) type of scholarship, with possible values A (assigned to excellent students), B (assigned to very good students), C (assigned to good students) and No (students without a scholarship); (c) class attendance level for each student, with possible values low, medium and high; and (d) like cars: a dichotomous variable randomly assigned to each student, with possible values Yes and No. This data set can be downloaded from http://u.pc.cd/JjCrtalK (accessed on 1 December 2022).
- **Zoo** data set was created by Richard Forsyt, and it is publicly available on https://archive.ics.uci.edu/ml/datasets/Zoo (accessed on 1 December 2022). The data set

has 16 features. Each sample is assigned to one out of seven classes or types. The name of the animals is also given.

- **Breast Cancer** data set. The samples are described by 9 attributes and 1 class, some of which are interval, and some of which are nominal. It can be downloaded from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer (accessed on 1 December 2022).

Table 4. Data sets used for the experiments.

Data Set	Total of Instances	Missing Values	Features		
			Nominal	*Numeric*	*Other*
BI Software recommendation	100	No	10	1	No
Scholarship	100	No	3	1	No
Zoo	101	No	1	2	Boolean (15)
Breast Cancer	286	Yes	6	1	Interval (3)

3.3. Data Preparation

The following general procedures were considered for the application of the proposed method:

1. Real valued variables are not analyzed; therefore they are identified automatically (in Python), and they are not processed.
2. In some data sets, categorical variables are encoded as integers or Boolean values. In these cases, a manual transformation of the values from integer type into categorical type was carried out.
3. Samples with missing values in one or more variables were removed from data sets.
4. Some variables do not provide important information, so they were removed manually. Examples of these variables are names, identifiers, consecutive numbering, etc.

4. Results

Tests of independence are commonly used to find possible associations between pairs of categorical variables. Our method allows detecting hidden associations between this type of variables, which otherwise would be very difficult to discover, even using the Chi-square test or Fisher's exact test manually. The implementation of the method includes the generation of a graph to facilitate the understanding of the results. These are shown in the next section.

4.1. BI Software Recommendation Data Set

This data set has 100 samples and 11 features. We get rid of the following variables: *product_id* (the identifier of the product), which is not significant; *rating*, which is a real number between one and five; *category*, which has 18 levels.

Figure 1 shows the number of associations detected by the method. The proposed method discovered that the highest number of associations in the BI Software recommendation data set occurred for the variables *industry* and *pricing*.

Table 5 shows one of the contingency tables in which an association is detected with Chi-square test. Variable *Industry* has 11 levels: manufacturing, fashion, utilities, marketing, consultancy telecommunications, IT pharma, food, academia and retail; whereas variable *Princing* has tree levels: (freemium, opensource and enterprise). Manually finding the combinations that are statistically relevant is a time-consuming task. Here is where the method is useful. The implementation of the method provides the results as a figure, the tables of contingency and the corresponding p-values.

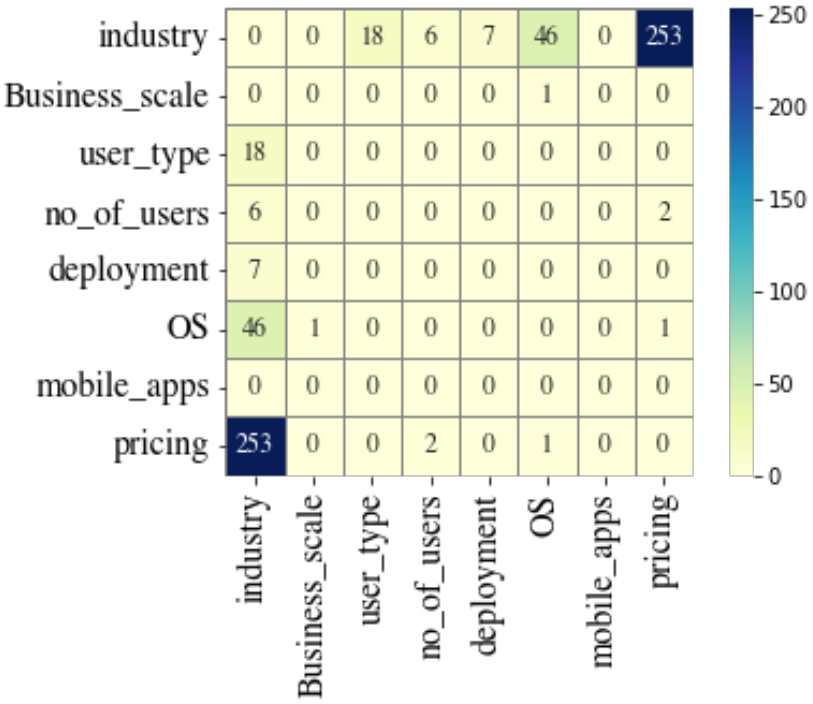

Figure 1. Associations detected in the BI Software recommendation data set.

Table 5. An example of the 253 relationships found between the industry and price variables. The contingency table is shown.

		Pricing	
		(Freemium)	**(Open Source, Enterprise)**
Industry	**(Manufacturing, Fashion, Utilities)**	15	17
	(Marketing, Consultancy Telecommunications, IT Pharma, Food, Academia, Retail)	17	51

p-value = 0.0002.

4.2. Scholarship Data Set

For the case of the Scholarship synthetic data set, the *score* was treated as a categorical variable.

In this data set, the method detects many associations between the variables *score*, *type of scholarship* and *class attendance level*, see Figure 2. It is important to remember that tests of independence only identify associations, not causation. However, for this data set, the results found are reasonable, since students with high grades have some type of scholarship and generally attend more classes than other students. On the other hand, liking cars has nothing to do with grades or scholarships. As the values of this variable were imputed randomly, the number of associations found is very low.

The values shown in the cells of Figure 2 indicate the number of associations detected between pairs of variables.

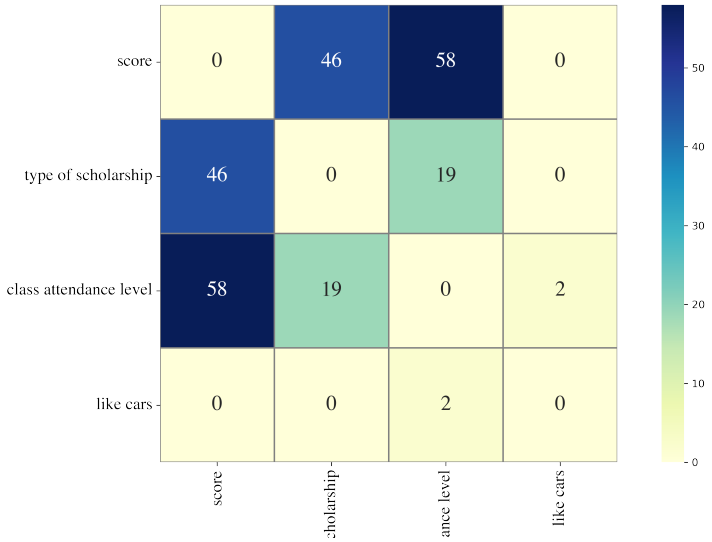

Figure 2. Associations detected in Scholarship data set.

4.3. Zoo Data Set

In this data set, all the Boolean values were treated as dichotomous (categorical). The variable *animal-name* was deleted manually. The variables *legs* (number of legs) and *type* (category of animal) were treated as categorical variables. Figure 3 shows the associations detected by the method. The *legs* and *type* variables are the ones with the greatest number of associations.

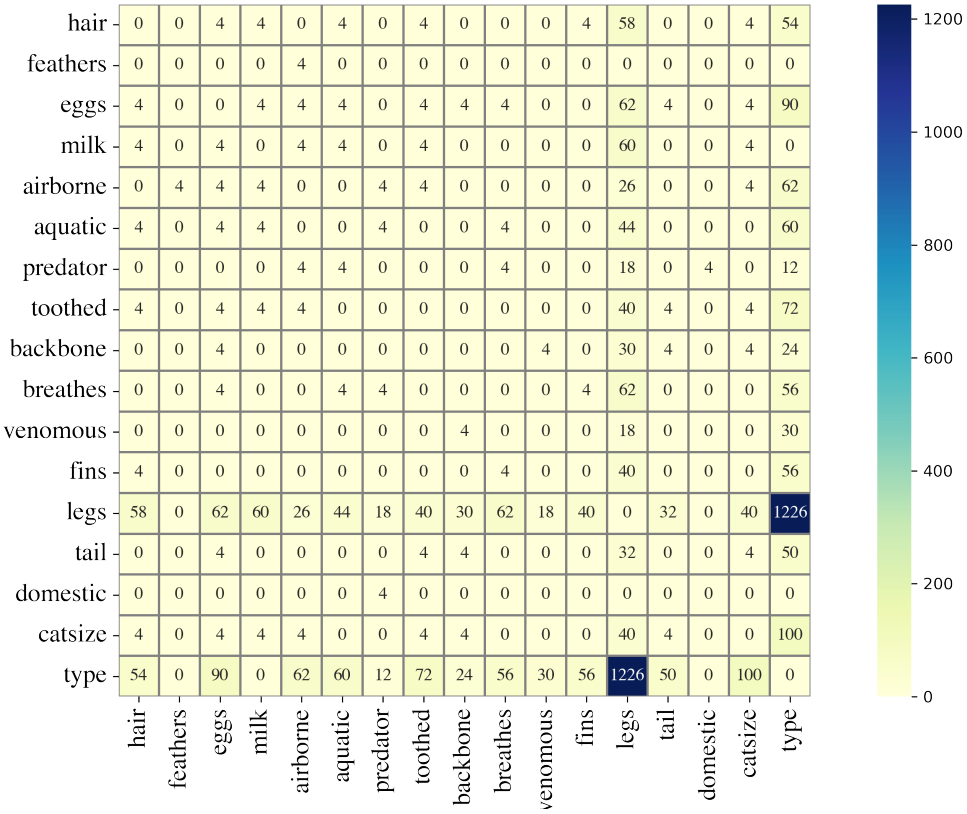

Figure 3. Associations detected in Zoo data set.

In this data set, we manually removed some variables to more clearly show the associations; these are presented in Figure 4. It can be seen that variable *feathers* has a relationship with variable *airborne*, which sounds logical. Variable *fins* has a relationship with variable *hair*. This could be because animals with fins do not have hair. A veterinarian or animal expert could explain the logic behind the associations detected by the method or whether they make any real sense. It is also possible that the method simply detected associations by the distribution of the data. In either case, the method presents the possible associations between variables.

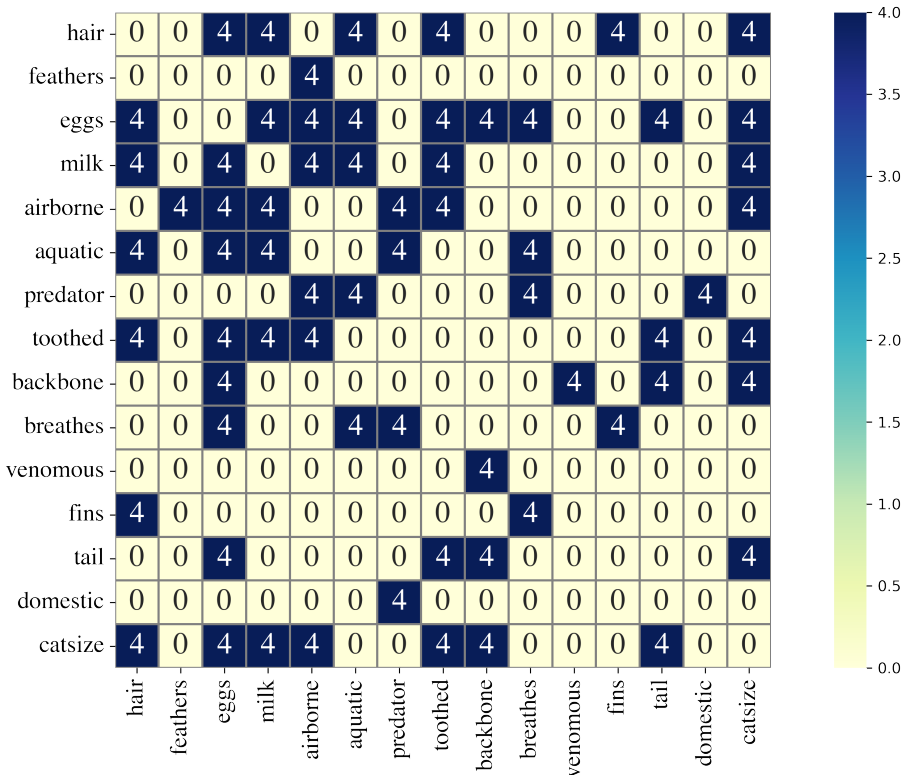

Figure 4. Results using only some variables of Zoo data set.

4.4. Breast Cancer Data Set

The Breast Cancer data set contains variables of type interval; these are *age*, *tumor-size* and *inv-nodes*; the variable *deg-malig* has values 1, 2 and 3. The rest of the variables are of type categorical. All non categorical features were transformed into categorical ones. Figure 5 shows the results obtained by the application of the method. It is noticeable that for this data set, the computation time was more than 7 min on a MacBook Pro, Apple M1 processor with 8 GB RAM.

In this data set, the variables with the highest number of associations are *inv-nodes* and *tumor-size*. Although the interpretations on the associations detected are out of the scope of the expertise of the authors, the information retrieved by the method could be useful for an oncologist.

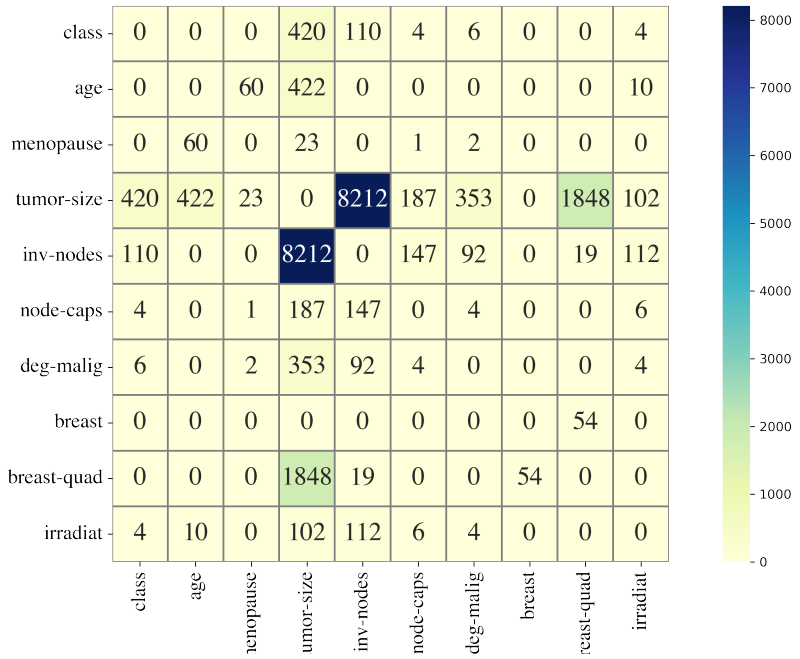

Figure 5. Associations detected in the Breast Cancer data set.

5. Conclusions

Independence tests are a very powerful tool for determining the association of two categorical variables. In the literature, there has been much discussion about the validity of these tests when one or more cells with low values or zeros are presented in the contingency tables. In this article, this problem is tackled from a computational point of view, proposing a method that minimizes the degree of freedom of the contingency tables by grouping the levels of each analyzed variable. In this way, the hidden association between the variables can be detected.

The implementation of the method in software generates a visual representation of the total associations found between the variables, allowing the expert to more easily identify if the association found has any substantial and real meaning.

To demonstrate the method, the synthetic data set "Scholarship" was generated in which some type of scholarship is assigned to students with high grades, and no type of scholarship to students with low grades. In addition, class attendance is higher in students with a scholarship than in the others. Furthermore, the set Scholarship has a variable whose values were generated pseudo-randomly. The proposed method finds a high number of associations between the variables *score*, *type of scholarship* and *class attendance level*, and two associations between the random variable, just as expected. The Scholarship data set, and the Python implementation of the proposed method are publicly available for download.

The method was also tested with real-world data sets, also publicly available on the Internet, finding a high number of associations between several variables.

The proposed solution allows identifying associations between multilevel categorical variables, performing an exhaustive search among all possible combinations of levels. The results are displayed graphically, facilitating the identification of the pairs of variables with the greatest number of possible associations. However, because many of these partitions might not be important, the researcher still has to manually analyze the generated crosstabs. This is a disadvantage of the method in which we currently work. Furthermore, we are improving the method in the computational part because we noticed that with variables of ten or more levels, the processing can take hours. An evolutionary computing approach to solve this problem is still under development.

Author Contributions: Conceptualization, A.L.-C., L.R.-M., F.G.-L. and M.Q.-L.; formal analysis, A.L.-C.; investigation, L.R.-M., F.G.-L., M.Q.-L. and C.A.R.-H.; methodology, A.L.-C., L.R.-M., F.G.-L., M.Q.-L. and C.A.R.-H.; resources, A.L.-C.; software, C.A.R.-H.; supervision, A.L.-C.; validation, A.L.-C., L.R.-M., F.G.-L. and M.Q.-L.; writing—original draft, A.L.-C., L.R.-M. and F.G.-L.; writing—review and editing, A.L.-C., L.R.-M., M.Q.-L. and C.A.R.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad Autónoma del Estado de México.

Data Availability Statement: Zoo and Breast Cancer data sets are publicly available at [21]. Scholarship data set was generated by the authors, it is available at http://u.pc.cd/JjCrtalK (accessed on 1 December 2022).

Acknowledgments: Authors thank to Universidad Autónoma del Estado de México for the support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fisher, M.J.; Marshall, A.P.; Mitchell, M. Testing differences in proportions. *Aust. Crit. Care* **2011**, *24*, 133–138. [CrossRef] [PubMed]
2. Olawande, T.I.; Okagbue, H.I.; Jegede, A.S.; Edewor, P.A.; Fasasi, L.T. Survey datasets on patterns of utilization of mental healthcare services among people living with mental illness. *Data Brief* **2018**, *19*, 2095–2103. [CrossRef] [PubMed]
3. Bustos, M.E.R.; Vaca, M.T.; de la Torre, C. Actividad física y adaptación escolar en estudiantes de medicina en un campus de la Universidad Nacional Autónoma de México. *Investig. Educ. MéDica* **2017**, *6*, 16–24. [CrossRef]
4. Targum, S.D.; Catania, C.J. Early treatment response affects signal detection in a placebo-controlled depression study. *Pers. Med. Psychiatry* **2017**, *4–6*, 19–24. [CrossRef]
5. Chen, Z.; Huang, H.; Ng, H.K.T. An improved robust association test for GWAS with multiple diseases. *Stat. Probab. Lett.* **2014**, *91*, 153–161. [CrossRef]
6. Adamu, P.I.; Oguntunde, P.E.; Okagbue, H.I.; Agboola, O.O. Statistical data analysis of cancer incidences in insurgency affected states in Nigeria. *Data Brief* **2018**, *18*, 2029–2046. [CrossRef] [PubMed]
7. Losappio, L.M.; Cappai, A.; Arcolaci, A.; Badiu, I.; Bonadonna, P.; Boni, E.; Bussolino, C.; Caminati, M.; Galati, P.; Heffler, E.; et al. Anxiety and Depression Effects During Drug Provocation Test. *J. Allergy Clin. Immunol. Pract.* **2018**, *6*, 1637–1641. [CrossRef] [PubMed]
8. Haberman, S.J. A Warning on the Use of Chi-Squared Statistics With Frequency Tables With Small Expected Cell Counts. *J. Am. Stat. Assoc.* **1988**, *83*, 555–560. [CrossRef]
9. Bradley, D.R.; Bradley, T.D.; McGrath, S.G.; Cutcomb, S.D. Type I error rate of the chi-square test in independence in R×C tables that have small expected frequencies. *Psychol. Bull.* **1979**, *86*, 1290–1297. [CrossRef]
10. Software, M. Minitab Support. Available online: https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/tables/supporting-topics/chi-square/are-the-results-of-my-chi-square-test-invalid/ (accessed on 1 December 2022).
11. Sharpe, D. Your chi-square test is statistically significant: Now what? *Pract. Assess. Res. Eval.* **2015**, *20*, 1–10.
12. Zeng, Y.; Chen, Y.; Yuan, Z. iSuc-ChiDT: A computational method for identifying succinylation sites using statistical difference table encoding and the chi-square decision table classifier. *BioData Min.* **2022**, *15*, 3. [CrossRef] [PubMed]
13. Overstall, A.M.; King, R. A default prior distribution for contingency tables with dependent factor levels. *Stat. Methodol.* **2014**, *16*, 90–99. [CrossRef] [PubMed]
14. Álvarez de Toledo, P.; Núñez, F.; Usabiaga, C. Matching and clustering in square contingency tables. Who matches with whom in the Spanish labour market. *Comput. Stat. Data Anal.* **2018**, *127*, 135–159. [CrossRef]
15. Lalanne, C.; Mesbah, M. 3-Measures and Tests of Association Between Two Variables. In *Biostatistics and Computer-Based Analysis of Health Data Using R*; Lalanne, C., Mesbah, M., Eds.; Elsevier: Amsterdam, The Netherlands, 2016; pp. 41–63. [CrossRef]
16. Pandis, N. The chi-square test. *Am. J. Orthod. Dentofac. Orthop.* **2016**, *150*, 898–899. [CrossRef] [PubMed]
17. Gilbert, G.E.; Prion, S. Making Sense of Methods and Measurement: The Chi-Square Test. *Clin. Simul. Nurs.* **2016**, *12*, 145–146. [CrossRef]
18. Parkinson, S.; Khan, S. Identifying irregularities in security event logs through an object-based Chi-squared test of independence. *J. Inf. Secur. Appl.* **2018**, *40*, 52–62. [CrossRef]
19. Daya, S. Fisher exact test. *Evid.-Based Obstet. Gynecol.* **2002**, *4*, 3–4. [CrossRef]
20. Hämäläinen, W. New upper bounds for tight and fast approximation of Fisher's exact test in dependency rule mining. *Comput. Stat. Data Anal.* **2016**, *93*, 469–482. [CrossRef]
21. Dheeru, D.; Karra Taniskidou, E. UCI Machine Learning Repository. Available online: http://archive.ics.uci.edu/ml (accessed on 1 Decemeber 2022).