VISUAL ANALYTICS IN HIGH-DIMENSIONAL DATA WITH DICHOTOMOUS OUTCOME

by

Chong Zhang

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2017

Approved by:

_____

Dr. Jing Yang

_____

Dr. Zbigniew Ras

_____

Dr. Zachary Wartell

_____

Dr. Bill Chu

_____

ABSTRACT

CHONG ZHANG. Visual analytics in high-dimensional data with dichotomous outcome. (Under the direction of DR. JING YANG)

High-dimensional data becomes common in application areas such as environmental studies and healthcare. The high dimensionality presents opportunities for understanding how certain outcomes happen by identifying significant variables contributing to the outcomes. Many efforts have been made to address this task. However, automated data analysis techniques often suffer from the "curse of dimensionality" and the difficulty of result interpretations. To integrate human intelligence into the analysis process and facilitate information communication with users, high-dimensional data visualization techniques have been developed. Unfortunately, high-dimensional data often leads to a cluttered visual display that obscures pattern discovery and hinders understanding of the data. Whereas a few visual analytics approaches have been developed to bridge automated data analysis and interactive visualization for high-dimensional data, few existing works have been focused on finding explanatory relationships between variables and outcomes.

In this dissertation, we address the task with two distinct paths from high-dimensional data with dichotomous outcomes to knowledge. First, we use visualizations to facilitate logit model building. We propose two approaches. In the first approach, Parallel Coordinates is used to facilitate dimension reduction based on correlation analysis, the first step of logit model building. It addresses the difficulties of correlation comparison and exploration when there are hierarchical outcome variables. In the second

approach, a visual analytics pipeline is proposed for logit modeling. It leverages the traditional modeling pipeline by providing (1) intuitive visualizations for inspecting statistical indicators and the relationships among the variables and (2) a seamless, effective dimension reduction pipeline for selecting variables for inclusion in high quality logistic regression models.

Second, we enhance visualizations with automated data analysis. In particular, association rule mining is employed to enhance Parallel Sets for categorical data exploration. Dimension reduction and reordering are conducted to reduce clutters and facilitate visual explorations in Parallel Sets based on significant association rules. The effectiveness and efficiency of our approaches are illustrated by a set of case studies and experiments with benchmark datasets.

ACKNOWLEDGMENTS

First of all, I am very grateful to my advisor, Dr. Jing Yang for advising me how to do research and for having an open mind for every idea that came along. Her encouragement and support helped me complete this dissertation. I appreciate the valuable suggestions from my committee members, Dr. Zbigniew Ras, Dr. Zachary Wartell, and Dr. Bill Chu.

I have had the great fortune to meet with Tamara Munzner and Daniel A. Keim for feedback on the design study of EPA research project. I would also like to thank Jun Li, Jiancheng Jiang, and Weihua Zhou for comments and suggestions on the statistical methodologies.

This work was mainly supported by the U.S. Environmental Protection Agency grant (#R834790). It was also partly supported through a cooperative agreement (U01dd000494) between the Centers for Disease Control and Prevention (CDC) and the Texas Department of State Health Services (TX DSHS). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the USEPA, the CDC, or the TX DSHS. Further, USEPA and the authors do not endorse the purchase of any commercial products or services mentioned in the publication.

I acknowledge the love and support of my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 High-Dimensional Data with Dichotomous Outcome

High-dimensional data is characterized by a larger number of dimensions than that considered in classical multivariate analysis. Dichotomous outcome has only two categories or levels for a particular dimension whose variation is being studied. For example, in a birth defect dataset, continuous explanatory variables are exposure to chemicals and the outcome is "yes" or "no" for having or not having a birth defect. In the Mushroom dataset [68], categorical explanatory variables are mushroom shape, color, odor, etc. and the outcome is "edible" or "poisonous".

High-dimensional data presents many challenges as well as opportunities. One of the problems with the high dimensionality is that not all observed variables are important for understanding the event/response/outcome of interest. It is of interest to study how the dichotomous outcome and explanatory variables are related. Identifying significant explanatory variables and examining them are critical tasks. They are the tasks addressed in the dissertation.

Regression models [53] can be used to help understand how the outcome variable changes when any one of the explanatory variables is varied. The interpretation of regression coefficients (β parameters) is the expected change in the outcome variable for a one-unit change in an explanatory variable while holding other explanatory

variables in the model constant. Since the outcome variable is dichotomous, logistic regression is used.

Ockham's razor (the principle of parsimony) [88] states that that among several plausible explanations for a phenomenon, the simplest is best. High-dimensional regression modeling seeks the most parsimonious model that still explains the data. The rationale is that the resultant model is expected to be statistically stable and be more easily generalized. The more variables included in a model, the more noises brought, and the greater estimated errors become [46, 29, 41].

Automated variable selections such as Stepwise Selection [26], Ridge Regression [18], LASSO [89], Elastic Net [101] often suffer from the "curse of dimensionality" and results in models that are unstable or under-specific (see more details in Chapter 3). To integrate human intelligence into the analysis process and facilitate information communication with users, high-dimensional data visualization techniques have been developed. Unfortunately, high-dimensional data often leads to a cluttered visual display that obscures pattern discovery and hinders understanding of the data. This is why Visual Analytics come into play which bridge automated data analysis and interactive visualization.

1.2 Visual Analytics

According to [87], Visual Analytics is "the science of analytical reasoning facilitated by interactive visual interfaces". Nowadays, more and more data is produced. The ability to store the data is increasing faster than the ability to analyze it. In recent decades, a lot of automated data analysis methods have been proposed in quantitative

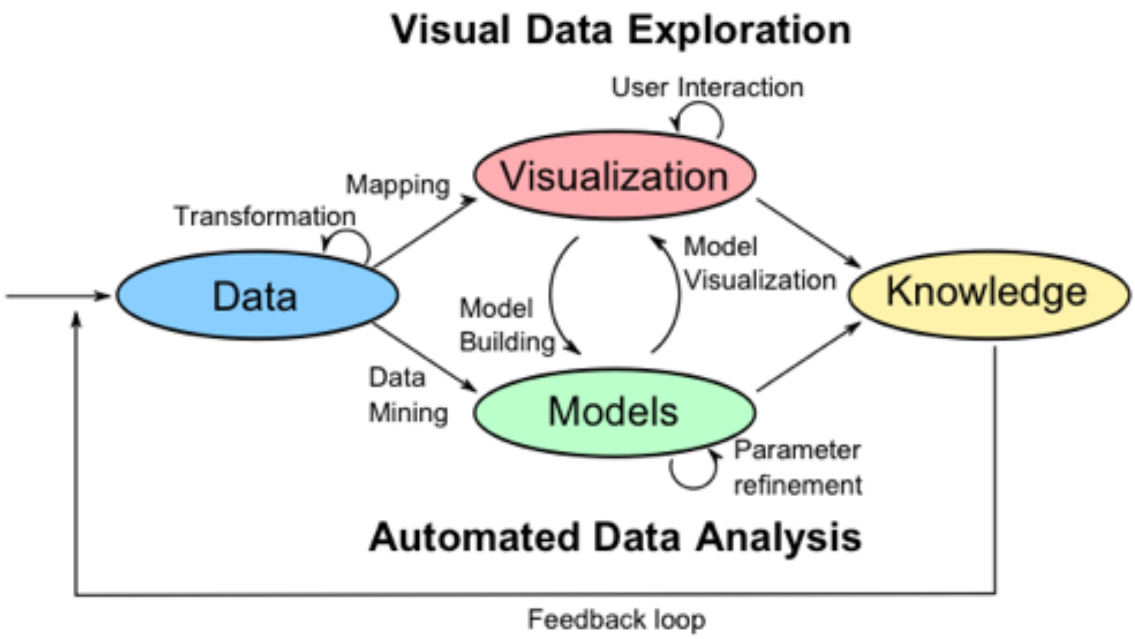

Figure 1: The Visual Analytics Process, adopted from Keim et al. [58, p. 10]

domains. However, human intelligence is necessary for some complex analyses. As Stuart G. Walesh 1989 said, "The computer is incredibly fast, accurate and stupid. Man is unbelievably slow, inaccurate and brilliant. The marriage of the two is a challenge and opportunity beyond imagination."

As shown in Figure 1, the Visual Analytics Process combines automated analyses and interactive visualizations to gain knowledge from data [58]. The first step is data transformation. After that, there are two directions that can be chosen in order: visual data exploration and automated data analysis. If an automated data analysis is chosen first, statistical and/or data mining methods are often applied to create models. Once a model is created, visualization then can be used to evaluate and refine the model by modifying parameters or changing algorithms. If visual data exploration is chosen first, findings in the visualization can be used to build models. In the two directions, interactions are always important in terms of discovering knowledge from

data.

We explore both of the two paths from data to knowledge as described in the Figure 1 for high-dimensional data with dichotomous outcome. We leverage the advantages of visualization and quantitative analysis to propose new visual analytic approaches to high-dimensional correlation analysis and logistic regression model building. We also discuss how association rule mining may create a new dimension ordering approach to reducing visual clutter. This dissertation illustrates the complements between visualization and quantitative analysis. On one hand, visualization make the intermediate or final quantitative analysis results easier to be understood. On the other hand, the quantitative results have a potential to change visuals for a better communication of knowledge discovering.

1.3 Dissertation Organization

Correlation analysis can be the first step of regression modeling. It can reduce a large number of dimensions to a few that are statistically significant correlated to the outcome variable. However, when the outcome variable has a hierarchy that can be factorized into more specific dichotomous variables, correlation coefficients of dimensions can vary in these new outcome variables. Then using correlation threshold to filter dimensions needs to be flexible, because a threshold working well for one specific outcome may fail for another one with a different correlation distribution. In chapter 2, we propose a visual analytic approach to filtering unrelated explanatory variables according to whether they are significantly correlated with the outcome variables. Parallel Coordinates [48] is used to facilitate correlation analysis for datasets with

hierarchical outcome variables where scientists can intuitively examine the difference of correlations with different outcomes.

The correlation analysis in chapter 2 also raised new questions, such as whether the correlated variables are risk factors, do they work as different groups, is there any third variables that affect the relationship, and would demographic characteristics change the relationship. To provide a deeply investigation about the sophisticated relationship between variables, in chapter 3, we propose a visual regression model building pipeline for datasets with a large number of explanatory variables. We leverage interactive visualizations to facilitate variable selection for logistic regression model building.

Because of the different nature of categorical variable, when the relationship between a dichotomous outcome and categorical variables is studied using regression analysis, the additional step of "dummy coding" is conducted [53, 3, 46]. However, the coding leads to a much more number of dimensions resulting in the increases of "curse of dimensionality". Next in chapter 4, in the context of high-dimensional categorical analysis, we explore how to take advantage of association rule mining to facilitate visual categorical data exploration of high-dimensional categorical datasets.

CHAPTER 2: PARALLEL COORDINATES TO FACILITATE CORRELATION ANALYSIS

In this chapter, we introduce an epidemiological application of parallel coordinates in high-dimensional correlation analysis.

2.1 Introduction

2.1.1 The Birth Defect Research Project

Birth defects are structural or chromosomal abnormalities that a baby has at birth. Despite the significant morbidity and mortality associated with these conditions, causes for an estimated 65% to 75% of birth defects remain unknown [15]. The United States Environmental Protection Agency (EPA), under the Toxics Release Inventory (TRI) Program, requires U.S. facilities to report yearly their release of more than 650 toxic chemicals into the environment.[1] Despite many studies, scientists do not fully understand the specific relationships between the environment and abnormalities. To help uncover these relationships, we have worked on a project with the aim of revealing associations between maternal exposure to air pollutants and congenital malformation in offspring.

The project was sponsored by the US EPA beginning in 2011. The research team includes researchers in geographic information science, computer science, and epidemiology. Data used in this project includes (1) TRI facility locations in Texas and

[1] https://www.epa.gov/toxics-release-inventory-tri-program/tri-listed-chemicals

toxic emissions from these locations from 1996 to 2008 and (2) birth defect case and control data obtained from the Texas Birth Defects Registry along with birth records.

2.1.2 Data Description

A large, high-dimensional environmental health dataset has been created for this study. In particular, information for 60,613 cases (births with major congenital malformations) and 244,927 controls (births without major congenital malformations) between 1996 and 2008 were retrieved from the Texas Birth Defects Registry and birth records. The dataset includes neural tube defects, heart defects, oral clefts, and limb reduction defects. Information about the release of 449 toxic chemicals in Texas during the same period was retrieved from the Toxics Release Inventory (TRI) database of the United States Environmental Protection Agency (EPA). The case and control data and the TRI data were linked using procedures developed by the research team [12, 98]. For each case and each control, maternal exposures to the 449 chemicals were calculated and recorded as numerical values [100]. Five maternal and infant characteristics, such as the mother's age group, level of education, and the gender of an infant, are recorded as categorical values. The response variable defines whether a child is born with one of the selected birth defects and the explanatory variables are the chemical exposure and the maternal and infant attributes.

The outcome variable is hierarchical. For example, neural tube defects has more specific defects such as anencephaly and spina bifida. Heart defects include conotruncal heart defects, septal heart defects, atrioventricular septal heart defects, obstructive heart defects, etc.

The data is high-dimensional, hierarchical, mixed continuous and categorical variables. Furthermore, maternal exposure to pollutants is often encoded as zeros in the dataset, resulting in sparse datasets that are difficult to explore efficiently.

2.1.3 Motivation

To identify risk factors, filtering unrelated independent variables according to whether they are strongly correlated to the dependent variables is effective, intuitive, and inexpensive compared with other dimension reduction methods. Therefore, before using any statistical models, analysts often conduct correlation analysis, which may remove a large number of unrelated independent variables from the subsequent regression analysis.

Rather than replacing the statistical analysis pipeline used by the researchers, the visual analytics approach follows their common practice. Visualization and statistical analysis are carefully inserted into the pipeline where statistical analyses face bottlenecks. For example, when researchers analyze global risk factors while ignoring characteristic differences, they usually first test the correlations and dependencies between the pollutants and the birth defects, only the variables passing the tests are then analyzed using a logit model [76]. The main bottleneck have been identified in this stage: the correlations among hundreds of pollutants and multiple types of birth defects overwhelm the researchers and it is difficult for them to answer questions using automated approaches:

1. What correlation threshold should be used for dimension reduction? The number of dimensions selected can vary widely with a small change in the threshold.

In addition, a threshold working well for one birth defect may fail for another one with a different correlation distribution.

2. Should a birth defect group be split? To make sure that there are enough cases, the researchers might group birth defects, such as all neural tube defects, instead of examining anencephaly and spina bifida as separate defects. However, birth defects in the same group may be associated with different risk factors. This can weaken the association observed and cause risk factor identification to fail. The distribution of the correlations between chemicals and a birth defect group may give an early warning of such situations.

3. Does a chemical have different correlations with different birth defects? Researchers are interested in chemicals with this feature since associations and possible risk factors may be missed by lumping similar birth defects together in this situation. The above tasks are complex with the number of chemicals and birth defects in this study.

To address this challenge, Parallel Coordinates [48] is used where researchers can intuitively examine the correlations and interactively set thresholds for individual birth defects with more confidence.

2.1.4 Case-Control Study

Case-control studies, retrospective in design, aim to identify possible factors affecting outcome and are useful for studying rare conditions [72]. It is the methodology used in this birth defect research project because birth defects are relatively rare

events, occurring in approximately 3% of live births. For a sample of subjects having Y = 1 (cases) and having Y = 0 (controls), the value of X is observed. Figure 2 provides an illustration of this dataset.

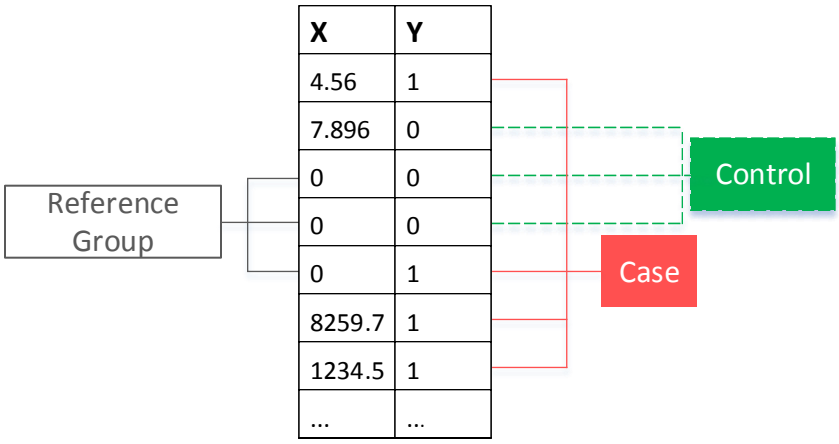

Figure 2: The Birth Defect Dataset observation example. A continuous dimension is denoted by X and a type of birth defect is denoted by Y. Case for this birth defect is recorded as 1 in the database while control is 0. Observations took 0 in X forms a reference group to conduct case-control study.

Evidence exists of an association between X and Y if the distribution of the X value differs between cases and controls [2]. For example, a 2 by 2 contingency table is a straightforward representation of such a comparison when X only has two levels (0 and 1), as shown in Table 1. Table cells display the count of subjects with the XY combinations.

Table 1: 2 X 2 contingency table

	Case	Control
Presence of exposure	a	b
Absence of exposure	c	d

A quantitative measure of the association between Y and X is **Odds Ratio (OR)**. It represents the odds that Y = 1 will occur given X = 1, compared to the odds of Y = 1 given X = 0. OR can be calculated using a 2 by 2 contingency table such as Table 1 with the following formula:

$$OR = ad/bc$$

OR can be used to determine whether X is a risk factor of Y as follows:

1. OR = 1: no difference when X = 0 or X = 1;
2. OR > 1: increased risk of Y = 1 when X = 1;
3. OR < 1: decreased risk of Y = 1 when X = 1.

Further, confidence intervals for ORs should be constructed to determine whether or not the association is statistically significant [24]. The confidence interval for an OR can be obtained from its natural log. The 95% confidence interval for odds ratio on the natural log scale is:

$$ln(OR) \pm 1.96 \times \sqrt{1/a + 1/b + 1/c + 1/d}$$

The exponential function is then used to get the lower and upper limits on the original scale of the odds ratio. If the 95% Confidence Interval does not contain the value of 1.0, the OR is statistically significant at the level of 0.05.

OR and logistic regression have a closed relationship. It can be derived from a logistic regression analysis result (see chapter 3.1.1). From the result, the regression coefficient for one explanatory variable is the difference in the log odds of the outcome

when there is one-unit increase in the value of the explanatory variable. In other words, the exponential function of the regression coefficient is the odds ratio associated with a one-unit increase in the explanatory variable.

2.2 Requirement analysis

In a high level, researchers need to identify the risk factors that contributed to the occurrence of certain types of birth defect. In order to make the goal more specific, requirement analysis was conducted by monthly phone conferences with the project research team supplemented by a face-to-face visit. The team members also communicated through written documents and live demos. From the meetings and visit, *Global Task* and *Local Task* for this project have been identified.

2.2.1 Global task

Researchers need to examine potential relationships between hundreds of chemicals and certain types of birth defects. The relationships might be assumed to be present in all mothers and infants with a given birth defect in this global task. Highly correlated chemicals can be analyzed as a whole. The birth defects are organized into groups (see Section 2.1.2 for detail). Birth defects in the same group can be studied as a whole or as specific defects to avoid missing risk factors of the individual defects.

The major challenges in this task is effectively and efficiently identifying potential patterns for detailed correlation analysis from many independent and dependent variables.

2.2.2 Local task

The cases and controls in the birth defect research project contain maternal and infant characteristics. Some of the characteristics associated with birth defects may not be direct causes of a birth defect, but modify the relation between an environmental exposure and risk of birth defects in offspring. Some mothers with these characteristics might have a higher risk for a given birth defect when exposed to certain toxic chemicals compared to mothers with the same exposure but without the characteristics. This phenomenon is known within epidemiological research as effect modification. For example, in a study on the relation between maternal residential proximity to Toxic Release Inventory industrial facilities and oral clefts in offspring [13], a maternal residence near such facilities, especially to facilities with heavy metals emissions, was more strongly associated with oral clefts among offspring of older women than younger women. Therefore, the local task has two goals: (1) to validate whether a population defined by the researchers is vulnerable and identify risk factors associated with it; (2) to thoroughly examine all the characteristics for vulnerable populations and their local risk factors. If there are many of them, the researchers are more interested in large populations and populations with higher risks.

Our approach is designed closely around the above tasks. It is introduced in the following sections.

2.3 Statistics Measures

Multiple statistical measures are used in our approach.

(1) Point Biserial Correlations (rpb): In my system, rpb [86] is used to measure the correlation between a birth defect (measured as a dichotomous variable) and a chemical exposure (measured as a continuous variable). Rpb is expressed in a range from +1 to -1; values larger/smaller than zero mean positive/negative correlation. The statistical significance of rpb needs to be tested and only correlations whose p-values are smaller than 0.05 are generally considered significant.

(2) Pearson's Chi-Square Test of Independence indicates whether an independent variable X can determine a dependent variable Y [64]. It can be used to determine whether there is a significant association between two categorical variables. This test is used in our approach to examine the independence between a birth defect and a categorical maternal or infant attribute.

(3) Crude Odds Ratios (OR) and their confidence intervals: OR is a quantitative measure of the association between a binary explanatory variable X and a binary outcome variable Y (see Section 2.1.4). OR and its 95% confidence interval are used for categorical variables [85]. Higher ORs indicate higher odds of the outcome. The association is significant if the lower/higher bound of the confidence interval is greater/smaller than 1 for a risk/protective factor. OR is crude in this step because it has not been adjusted for other variables.

2.4 Application of Parallel Coordinates for the Global Task

Our first step is to reduce the number of variables under consideration by removing those not having a significant correlation (see examples in [90] and [76]). Since the status of a birth defect is binary (yes or no) and the maternal exposure to a

pollutant is continuous, Point-Biserial Correlation (rpb) analysis [86] is used for correlation analysis. The resulting correlations between the birth defects and chemicals are visually presented to users in Parallel Coordinates [48] for browsing and dimension selection. This view allows users to intuitively browse the relationships between multiple birth defects or defect groups and a larger number of chemicals. According to the observation, the users can decide whether a group of birth defects can be studied as a whole or should be examined in more detail for better results. They can also interactively select an interesting birth defect and chemicals significantly correlated to it for further analysis.

Parallel Coordinates [48] is used to visualize the correlations and select independent variables significantly correlated to the birth defects. In Figure 3, each dimension represents a birth defect and each polyline represents a chemical. The position of the line on an axis encodes the rpbs between the chemical and the birth defect. Each rpb have a p-value indicating its statistical significance. A p-value threshold can be interactively set by the users and only correlations with a p-value below this threshold should be considered. Since chemical variables with high p-values are actually noise, we set them to zero to avoid distracting the users. Alternatively, it could be removed from the display, but this option is not used since removal causes a discontinued polyline (the correlation between the chemical and other birth defects may be significant).

From this view, users can examine the distribution of the rpbs on multiple birth defects. For example, Figure 3 shows that rpbs on Neural tube defects, Heart defects, and Limb reduction defects are well distributed, but the rpbs on Oral clefts are not.

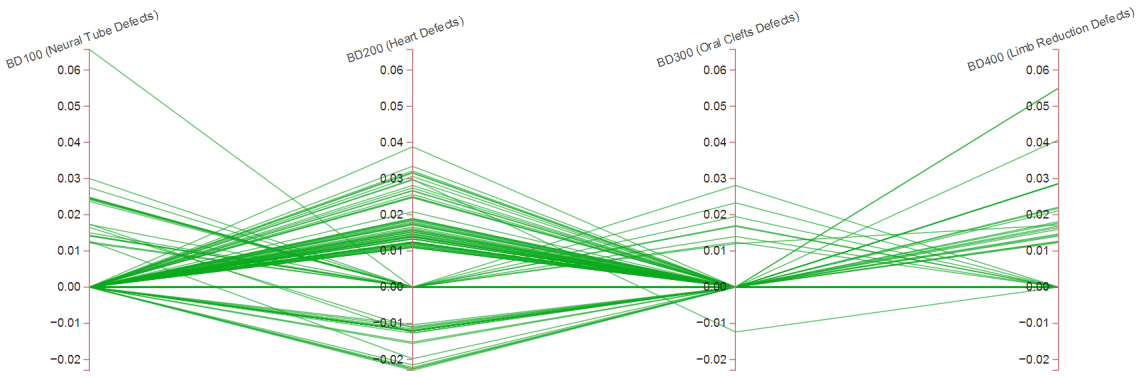

Figure 3: Correlation analysis with Parallel Coordinates. Each line represents a chemical and each axis represent a birth defect or a birth defect group.

This suggests that Oral clefts may need to be further divided into smaller categories (this was later confirmed since no risk factors were detected from the Oral clefts case and control group). Figure 4 shows the results when Oral clefts are divided into two groups, namely Cleft palate alone and Cleft lip with/without cleft palate. The rpb distributions on axes of Parallel Coordinates suggest that the Cleft palate signals are strong enough but the Cleft lip with/without cleft palate may need to be further split. When users move the mouse over a polyline, it will be highlighted and the name of the chemical will be displayed. In this way the correlation between the chemical and different birth defects can be examined. For example, the chemical *A1344281* (*ALUMINUM OXIDE (FIBROUS FORMS)*) is highlighted in purple in Figure 5. A 1-D brush is provided whose boundary defines a rpb range. Users can click an axis to trigger the brush and interactively change its boundary to select chemicals whose rpbs to the birth defect fall within that range and below the p value threshold. Selected chemicals are highlighted and their names are displayed. The total number of selected chemicals is also displayed (see Figure 5). This provides the users an instant visual feedback to the rpb threshold they set through the brush.

Figure 4: Correlation Analysis with Parallel Coordinates for Specific Types of Oral clefts

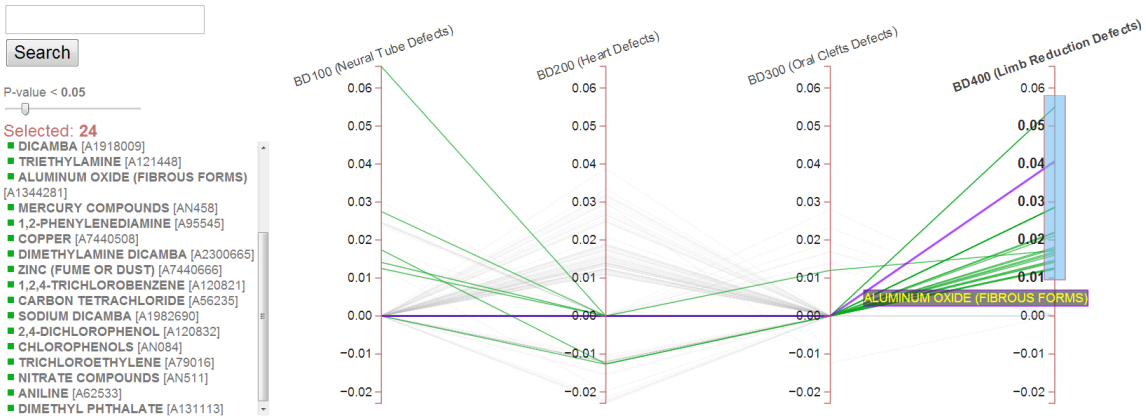

Figure 5: Brushing and Highlighting on Parallel Coordinates

2.5 Visual Analytics Approaches to the Local Task

A domain knowledge guided visual analysis is proposed to support the local tasks. In this approach, users can interactively select a maternal or infant characteristic to divide the data set into several sections. **Rose Graph**, a variation of the mosaic plot [34], is proposed to allow users to easily identify populations with high risks. Then, **Dual Axes Parallel Coordinates (DAPC)**, a novel variation of Parallel Coordinates [48], is used to visually compare the birth defect-pollutant correlations

within this population to those in the whole population. Chemicals with significantly increased correlations can be easily identified and selected for further analysis. The Rose Graph and the DAPC are coupled so that users can quickly examine varying populations.

2.5.1 Rose graph

The Rose graphs allows the analysts to examine the ORs of each population for multiple birth defects and select a population of interest (see Figure 6).

The ORs of a population are calculated using Equation 2.1.4. For calculating the OR of the population $M\_AGEG\_V = 6$, a is the number of cases where the age is greater than or equal to 40 and less than 100, b is the number of controls in the previous group, c is the number of cases where the age is less than 40, and d is the number of controls in the previous group. A Rose graph is drawn for each birth defect, whose name is drawn in the center of the graph. It consists of multiple sectors, each of which represents a population. The angle of a sector is proportional to the number of cases in this group, making it possible to observe both relatively big and small groups. The color and radius of a sector is mapped to the OR of the group. Populations with high ORs are vulnerable populations. They seem long and red in the graph and can be easily identified from the Rose graph. The dependency test is conducted between the variable and the birth defect. The result is drawn in the center of the graph beneath the name of the birth defect and can be referenced by the analysts later. Users can click a sector to examine the population it represents in the Dual Axes Parallel Coordinates.

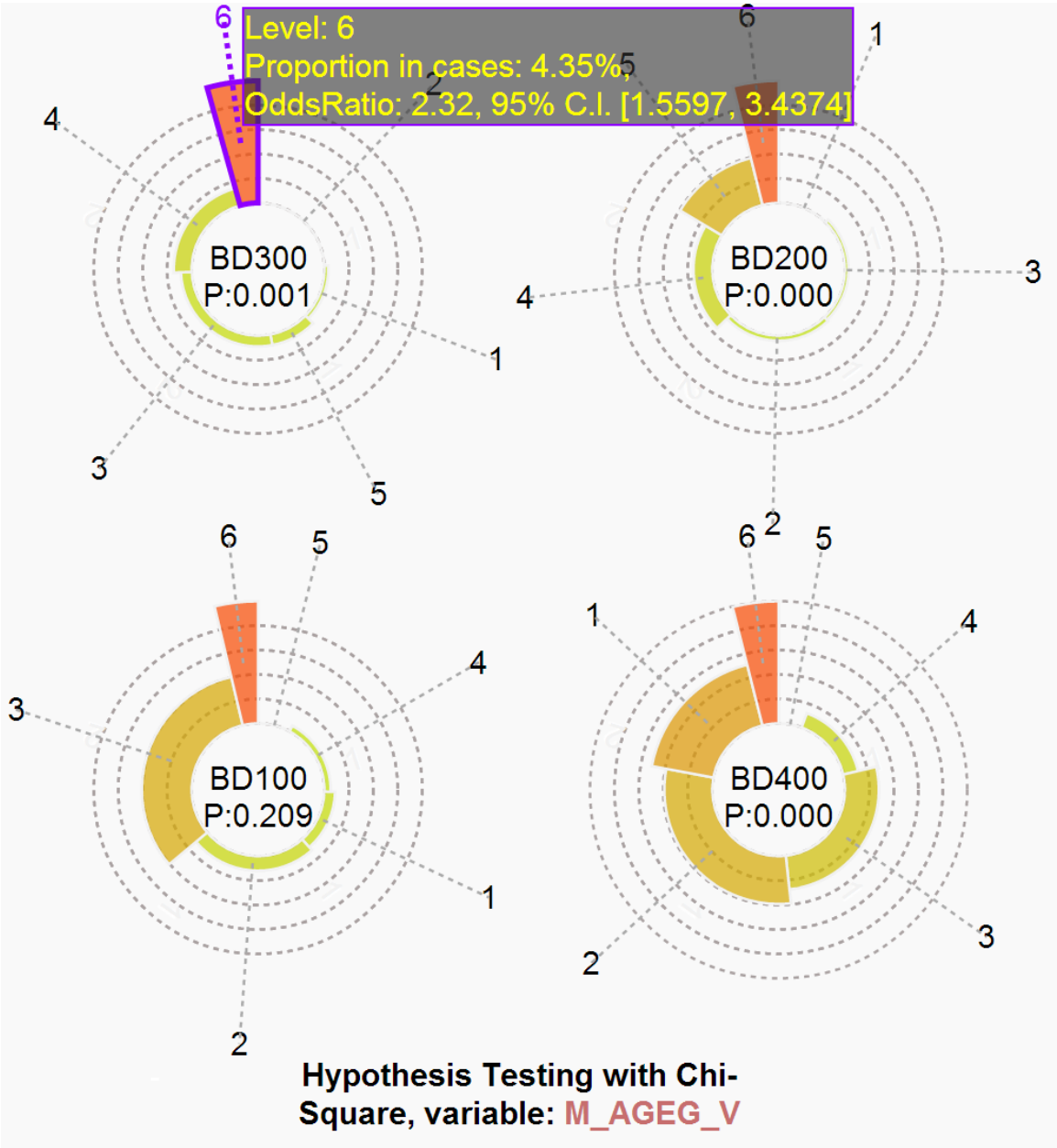

Figure 6: RoseGraph for demographic attribute `M_AGEG_V`. It has 6 categories. The variable is not statistical significant to BD100 because of the P-value of 0.209.

2.5.2 Dual Axes Parallel Coordinates (DAPC)

DAPC allows the analysts to intuitively compare the rpbs of the chemicals within this group and the global rpbs (see Figure 7).

In particular, a within group rpb is the rpb between a chemical and a birth defect

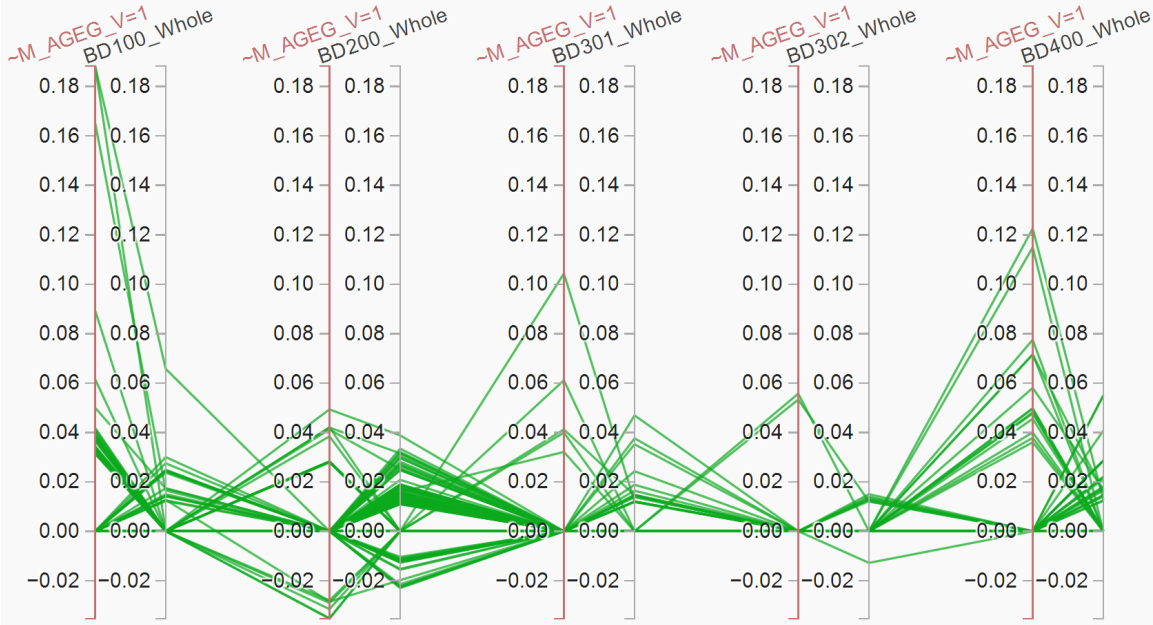

Figure 7: Correlation comparison between M_AGEG_V=1 and the entire dataset with the Dual Axes Parallel Coordinates.

calculated using the cases and controls of this group. A global rpb is the rpb between a chemical and a birth defect calculated using all the cases and controls of the birth defect, namely the rpb displayed using PC in the global analysis. In DAPC, each chemical is still represented by a polyline. There are two adjacent axes, in different colors, for the same birth defect. The within group rpbs between the chemicals and the birth defect are drawn on the left axis. The global rpbs between the chemicals and the birth defect are drawn on the right axis. It is easy to compare them for insights. For example, Figure 8 reveals that mothers with $M\_AGEG\_V = 6$ ($\geq$ 40 years old) are more vulnerable to all birth defects than other mothers when exposed to some chemicals. The analysts can interactively select chemicals from this view similar to the way selection is performed in the Parallel Coordinates used in the global analysis. They can also visually examine the relationship between a chemical and a birth defect

within the population from the pixel oriented view beneath the DAPC.

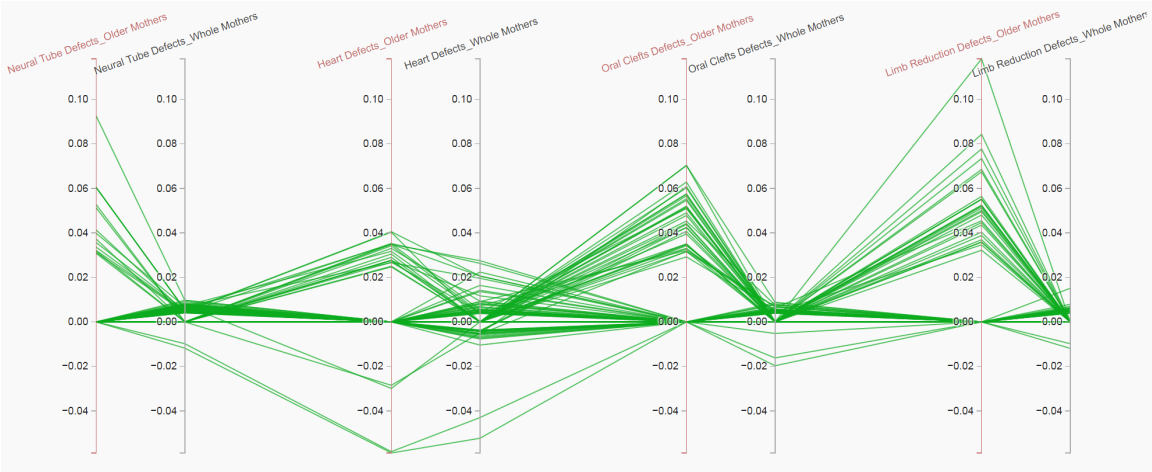

Figure 8: Correlation Comparison between M_AGEG_V=6 and the entire dataset. It shows that this population is more vulnerable to all birth defects than other mothers when exposed to some chemicals.

Using the Rose graph and the DAPC, researchers can easily identify vulnerable groups and potential risk factors of these groups. Then, the analysis can follow the steps in the global approach to identify chemicals that are risk factors of a birth defect. Adjusted ORs can be compared with results from the global analysis to learn if these chemicals are more dangerous to the vulnerable groups studied.

2.6 Discussion

Although much literature on multidimensional data visualization exists, there is still a wide gap between the requirements of practical multivariate applications and existing visualization techniques. For example, it seems that multivariate analysis in case-control studies has never been adequately supported with visualization approaches. Future multidimensional visualization studies should pay more attention to address the real needs of practitioners.

At the beginning of the project, we were searching for existing visualization tech-

niques, even the ones that seem very scalable and powerful, cannot provide a suitable solution to the problems facing the domain experts. Our attempt to design new visualization techniques to replace the statistical analysis methods in use was not accepted by the domain experts. However, when we use existing visualization techniques to complement the statistical analysis pipeline familiar to the experts, the experts immediately realized the value of the approach and started using it. This is not only because the experts can accept familiar approaches easier, but also because visual and statistical analysis have complementary advantages and disadvantages: visualizations are intuitive, allow user participation, and can provide users rich details of the data. However, visualizations generally lack the power to generate reliable quantitative results and their theoretical foundation is less matured than statistical analysis. Statistical analysis alone prohibits analysts from leveraging domain knowledge and perceptual capabilities to guide interactive exploration of their data. Since these features are quite complementary, a visual analytics approach that tightly integrates them can be very powerful.

2.7 Conclusion

In this chapter, we propose a solid visual analytics approach for case-control studies in large scale birth defect research. This approach is rooted in the common practices of domain experts and greatly enhances current scalability and effectiveness by providing intuitive correlation exploration visualization. It is among the earliest visual analytics techniques that support large scale case-control studies.

CHAPTER 3: VISUAL ANALYTICS IN LOGISTIC REGRESSION MODELING

3.1 Introduction

The approaches in chapter 3 showed that we can reduce the large number of chemicals to a few that are most correlated with certain types of birth defects, but these approaches are just a first step, and they raised several new questions:

- Since many variables in the dataset are interrelated with each other, whether the selected variable in the previous approach is a confounder or a risk factor? If it is a confounder, what risk factors could cause that? And if it is a risk factor, could there be any related confounders?
- If several numerical variables are selected for further analysis, how characteristic variable(s) modify the effects of selected numerical variables?

3.1.1 Logistic Regression

After the first step, typically there are still many pollutant variables to be further analyzed using regression analysis. Regression models [53] can be used to establish relationships between response variables and explanatory variables. The interpretation of regression coefficients (β parameters) is the expected change in the response variable for a one-unit change in an explanatory variable while holding other explanatory variables in the model constant. In our study, the response variable defines whether a child is born with one of the selected birth defects and the explanatory variables are

the chemical exposure and the maternal and infant attributes. Because the response variable is dichotomous, we model the logit-transformed probability [46] of having the birth defect $\pi(x)$ as a linear relationship with the explanatory variables. Univariate logit models,

$$logit[\pi(x)] = \beta_0 + \beta_1 X$$

, may lead to an abundance of false-positives [14, 85, 93] because the effect of X might be caused by other explanatory variables that are not considered in the model. In contrast, multivariate logit models,

$$logit[\pi(x)] = \beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_p Xp$$

, simultaneously measure the relationship between the dichotomous outcome variable and multiple explanatory variables. This allows the model to distinguish false-positives from true risk factors that should be further confirmed/rejected through subsequent epidemiological analyses.

3.1.2 Explanatory Modeling vs. Predictive Modeling

According to Shumeli [83], explanatory modeling is fundamentally different from predictive modeling. The former aims at summarizing and explanation, while the latter aims at empirical prediction. Explanatory modeling is retrospective, in that we want to discover the underlying relationships between the response variable and the explanatory variables after noise has been accounted for. When seeking an explanatory answer, the primary focus is on the data we have. For example, we want to know if it is true that exercising regularly (say 30 minutes per day) leads to lower

blood pressure. To answer this question, we need to collect data from patients about their exercise logs and their blood pressure values over time. The goal is to see if we can explain variations in blood pressure by variations in exercise. Blood pressure is impacted by not only exercise, but also a variety of other factors such as amount of intake sodium per day. These other factors would be considered noise in the above example as the focus is on the relationship between exercise and blood pressure.

Predictive modeling is forward-looking, in that we want to predict new observations using the known relationships between the data we have at hand. The known relationship may emerge from an explanatory analysis or some other techniques. For example, if I exercise one hour per day to what extent is my blood pressure likely to drop? To answer this question, we use a previously uncovered relationship between blood pressure and exercise to perform the prediction. In the above context, the focus is not on explanation, although an explanatory model can help with the prediction.

They are also different in the bias-variance. Bias is an error taken as the difference between the expected value of the model and the correct value which I'm trying to predict. Variance is an error taken as the variability of a model prediction for a given data point. The variance is how much the predictions for a given point vary between different training datasets. The shrinkage regression modeling such as ridge regression [18] and LASSO [89] shrink predictor coefficients or even eliminate them. They are very useful for prediction but not for explanation because they reduce variance by introducing bias.

In explanatory modeling, the focus is on minimizing bias to obtain the most accurate representation of the underlying theory. In contrast, predictive modeling seeks

to minimize the combination of bias and estimation variance, paying less attention to theoretical accuracy for improved empirical precision. Therefore, "wrong" explanatory models can sometimes have better prediction precision.

The criteria for variable selection differs significantly in these two contexts [83]. Explanatory modeling requires interpretable statistical models. Therefore, uninterpretable variable selection methods such as neural networks and k-nearest-neighbors are considered inappropriate. In predictive modeling, the top priority is the prediction accuracy. The variable selection methods can be interpretable or uninterpretable.

In the recent years, more and more visualization researchers have proposed interactive feature selection using visual analytics [94, 73, 10, 75, 62]. Visual analytics presents the aid in better understanding the problems that happened in variable selection and regression modeling. Despite all this, few targets explanatory regression model building for high-dimensional data with dichotomous outcome. Considering the associations between variables, we explore how visual analytics help address the problems that happened in explanatory regression modeling, such as multicollinearity, confounding, and weak effect.

3.1.3 Challenges

The high-dimensional nature of the dataset brings significant challenges to logit modeling. They include (1) **Overfitting**: A high-dimensional logit model may describe noise or random error instead of the underlying relationship between the outcome and explanatory variables [38, 46], (2) **Confounding**: A confounder is an extraneous variable associated with both the outcome and one or more true risk factors.

Along with the explanatory variable, it may explain all or part of the observed effect of the true risk factors thereby complicating and perhaps masking the true relationship between the outcome and the explanatory variables. Extreme confounding will lead to multicollinearity (see below), (3) **Multicollinearity**: If two highly correlated variables are placed into the same model, it may become unstable or over/underestimate variable effects [42, 17, 20], and (4) **Weak effect**: A variable might have a weak association with the outcome which may not be easily found without eliminating the influence of other related explanatory variables on the outcome. Because such associations are masked due to the existence of effects from the related variables, we need to include one of them at a time so that the weak effect can be seen compared to the unexplained variability.

Automatic approaches to selecting variables for logit model building, such as Stepwise selection [26], Ridge Regression [18], LASSO [89], and the Elastic Net [101], often result in models that are unstable, non-reproducible, or have extra parameters and bias. Therefore, it is desired to allow users to participate in explanatory model building. They can identify confounders, determine variable inclusion and exclusion, and interpret weak associations.

3.1.4 Contributions

To address this need, we propose a visual analytics approach that facilitates the building of high quality logit models for risk factor identification. To the best of our knowledge, this approach is among the first visual analytics efforts toward this purpose. The approach is general enough to be used in other application domains where

high-dimensional logit modeling is needed for explanation. The main contributions in this stage include:

- A design study where visual analytics techniques are developed for identifying birth defect risk factors from a high-dimensional environmental health dataset,
- A novel visual analytics approach to high-dimensional logit modeling. It seamlessly integrates statistical procedures and visualization techniques to make the modeling process easy, intuitive, and more accurate than automatic approaches. A fully working prototype has been developed which received positive feedback from two epidemiologists and a statistician,
- Case studies where the prototype was used to identify potential risk factors for limb reduction defects and characterize caravan insurance policy holders.

3.2 Related Work

3.2.1 Dimension Reduction

Dimension reduction or dimensionality reduction refers to the process of converting n dimensions of data set to k dimensions ($k < n$). These k dimensions can describe most of the variance within the original data set. There are a variety of techniques for doing this. Overall, these dimensions can be directly identified (filtered) or can be a combination of dimensions or a new set of dimensions that represent existing dimensions well.

Dimension reduction is beneficial to multiple aspects. It fastens computing by using lower number of dimension. It helps in removing redundant variables and reducing

multicollinearity, and as a result of that model performance can be improved.

3.2.1.1 Statistics and Data Mining

Principle Components Analysis (PCA) [56] is widely used for multivariate continuous variable analysis. It aims at reducing the dimension of the data by applying eigenvalue decomposition of data covariance or a correlation matrix. An orthogonal transformation is used to extract an internal linear combination of possibly correlated variables into a few that are uncorrelated groups, called principle components (PCs). The first PC is the linear combination with the largest variance (that is, accounts for as much of the variability in the data as possible). The second PC is the linear combination with the second largest variance and orthogonal to the first PC, and so on. The number of PCs is less than or equal to the number of original variables.

Factor analysis is closely related to PCA [55]. It seeks to describe the variability in the observed and correlated variables using a small number of underlying unobservable (latent) called factors. The idea behind factor analysis is that observed variables have similar patterns of responses because of their association with an underlying latent variable. The eigenvalue is a measure of how much of the variance of the observed variables a factor explains. It helps to decide how many factors to retain. The relationship of each variable to the underlying factor is expressed by the factor loading. Factor loadings can be interpreted like standardized regression coefficients. It decides which factors a variable belong to.

Factor analysis is a method of data reduction. It can be used to group interdependent variables into descriptive categories, groups, or clusters. It rotates factor axes

in an attempt to obtain simple and interpretable structures. The simple structures are patterns of results such that each original variable loads highly onto one and only one factor. Rotation methods are either orthogonal or oblique. Orthogonal rotation methods assumes that the factors are uncorrelated and impose a restriction such as varimax, equalmax, orthomax, quartimax, and promax [1].

Multidimensional Scaling (MDS) [63] uses pairwise dissimilarities to construct a map from the original high dimensional space to a lower dimensional space, preserving pairwise distances.

Dimension reduction has benefited from a great deal of work in both the statistics and data mining communities. I have limited the scope largely to the classical methods. There are other interesting methods as described in the [32].

3.2.1.2 Visualization

Automatic dimension reduction methods such as PCA [56] and MDS [63], are commonly used but produce results that had little user influence and are hard to interpret. To address this problem, many efforts have been made in visual dimension reduction by visualization researchers. Yang et. al. [96] propose organizing dimensions into a hierarchy according to their correlations and then visually presenting the hierarchy to users. Users can interactively navigate the tree and select nodes from the tree for interactive dimension reduction. Johansson and Johansson [52] present an interactive system for dimension reduction, where user-defined quality metrics are combined with weight functions to preserve important structures in the data. Fernstad et al. [31] use a set of interestingness measures for dimension filtering and organize correlated vari-

ables into clusters for effective subspace visual exploration. Paiva et al. [77] propose semi-supervised dimension reduction based on Partial Least Squares, where users can enhance the precision of the dimension reduction using their domain knowledge of the training set. DimStiller [47] is a visualization system with a dimension analysis and reduction workflow, where users can interactively transform the data using a variety of techniques. Our approach is for a specific statistical analysis task seldom addressed in visualization. It is unique in that its visual dimension reduction approaches are tightly integrated with an analysis pipeline familiar to analysts, allowing them to address important tasks once too difficult using only pure statistical methods.

3.2.2 Variable Selection for Regression Modeling

Regression modeling analyzes how the outcome variable is influenced by the explanatory variables. When a high-dimensional data is involved, variable selection is the necessary step before modeling. It is a process of selecting a subset of variables/features for use in model building. There are several reasons why it is needed:

- Irrelevant variables should be removed. They will add noise and blur the quantities of other variables that we are interested in [29, 41]. Ockham's razor (the principle pf parsimony) [88] states that that among several plausible explanations for a phenomenon, the simplest is best. Applied to model building, this implies that the smallest model that fits the data is best.

- Multicollinearity should be detected. Multicollinearity is a phenomenon where two or more variables in a regression model are highly correlated. It is often caused by having too many variables trying to carry the same information. The

regression model may become unstable or over/under-estimate variable effects [17, 46]

- Although small models are desired, the explanation should not be under-specified. If the model is missing one or more important explanatory variables, the model yields biased coefficients and biased explanation of the response.

In the field of statistics and data mining, variable selection is a classical topic. Stepwise selection [26] includes Backward Elimination, Forward Selection, and Stepwise Regression. Backward Elimination starts with all the explanatory variables in the model, removes the explanatory variable with highest p-value greater than a threshold, fits the model again, and then moves on to the second step. It stops when all p-values are less than the threshold. Forward Selection is the reversed version of the Backward Elimination. In particular, it starts with no variables in the model, checks p-values if the variables not in the model are added, chooses the variable with lowest p-value less than a threshold, and continues until no new explanatory variables can be added. Stepwise Regression is a combination of Backward Elimination and Forward Selection. This addresses the situation where variables are added or removed early and allows me to remove or add them back later. All the three Stepwise selection methods often exhibit high variance and severe bias [46, 89].

Ridge Regression [18] can be used to select variables that suffer from multicollinearity. In ridge regression, the first step is to standardize the variables (both response and explanatory variables) by subtracting their means and dividing by their standard deviations. All calculations are based on standardized variables. When the final

regression coefficients are received, they are adjusted back into their original scale. Ridge regression improves prediction error by shrinking large regression coefficients. However, in feature selection we cant simply choose features with the largest coefficients in the ridge solution. It also shrinks the coefficient estimates of all variables, and some researchers cannot accept the idea of restrictions on the betas. Besides, ridge regression adds a small value, k, to the diagonal elements of the correlation matrix. One can believe that the additional parameter k is essentially equivalent to largest beta.

LASSO [89] is also a shrinkage method. It minimizes the residual sum of squares (RSS) but poses a constraint to the sum of the absolute values of the coefficients being less than a constant. It is limited when handling highly correlated variables because it tends to choose only one among a group of variables with high correlations. The Elastic Net [101] suffers from a double amount of shrinkage which introduces unnecessary bias. In addition, these methods "shrink" the regression coefficients or eliminate them, which is not acceptable to domain experts such as epidemiologists or sociologists [83]. My visual analytics approach proposes a new solution to address this problem. The major statistics method used in my approach is regression models.

The automated feature selection methods described above often suffer from the difficulty of result interpretations. Human intelligence is not able to well-integrate the selection process.

In the visualization community, variable selection has been widely studied. The Value and Relation display [94] visually conveys the correlation among the variables with textures and distances and allows users to interactively select correlated or non-

correlated variables. DimStiller [47] is a visualization system with a dimension analysis and reduction workflow where users can interactively transform the data using a variety of techniques. SmartStripes [73] allows users to step through the feature selection process manually. Similar to my work, it uses feature partitions for dimension reduction. It is designed to be a preliminary analysis tool and does not consider cause-and-effect relationships. Fernstad et al. [31] use a set of interestingness measures for dimension filtering and organize correlated variables into clusters for effective subspace visual exploration. These works do not consider the association among explanatory and response variables in the regression relationship.

A few visual analytics approaches have been proposed for regression modeling in recent years. Steed et al. [84] use parallel coordinates to build linear regression models for hurricane activity prediction. Bögl et al. [10] apply line, bar, and scatter plots to a well-known statistical Box-Jenkins methodology for time series data regression modeling. Krause et al. [62] present a tool called "INFUSE" for comparing predictive features across feature selection and classification algorithms. Mühlbacher et al. [75] propose a partition-based framework for predictive linear regression model building. There are several differences between these approaches and my own. First, I focus on the visual analytics of multicollinearity, confounding, and weak effect instead of partitions. Second, many measures (i.e. R-squared, RMSE, and OLS) are not as well suited as logit regression because my response variables are not continuous. Third, all of those approaches target predictive modeling while explanatory regression modeling is needed in my project. According to Shumeli [83], explanatory regression modeling is fundamentally different from predictive regression modeling and this distinguishes my

work from existing efforts. Predictive regression modeling aims to minimize prediction errors but explanatory regression modeling emphasizes characteristic expressions and the relationships among variables for distinguishing between false-positive variables and true risk factors. The criteria for variable selection differ significantly in these two contexts [83].

3.3 Requirement Analysis

As the previous approach of correlation-based visual analytics has applied, the team members have characterized a set of comprehensive tasks through intensive meetings and discussions.

- **Task 1: Dimension Reduction.** According to a general guideline in logistic regression modeling [46], the complexity of multivariate studies can be reduced by first using univariate statistical indicators to filter out irrelevant or non-significant variables. These variables are ruled out as risk factors and confounders and do not need to be considered in further analysis. This requirement has been discussed in the chapter 3, but the difference is we bring multiple statistical indicators other than rpb for considering different aspects of measurements in this dimension reduction process.

- **Task 2: Dimension Relationship Analysis for Model Building.** Since univariate analysis may introduce false-positives, a variable should always be analyzed with other correlated variables in a multivariate logit model. To avoid problems such as overfitting, confounding, and multicollinearity, the relationship among the variables needs to be carefully examined. First, variables correlated

with the selected variables may need to be included into the model even if they are not in the initial candidate variable set. Second, groups of highly correlated variables need to be identified whose variability can be defined with a smaller set of latent variables. Variables within such groups should not be placed into a single super model. Rather, they should be placed into different models that are smaller and more stable [40]. Third, confounders need to be separated for further epidemiological analyses.

- **Task 3: Effect Change with Demographic Characteristics.** The cases and controls in this project contain maternal and infant characteristics. Some of the characteristics may not be direct causes of birth defects, but instead modify the relationship between environmental exposure and the risk of birth defects in offspring. To assess the effect of a chemical, demographic characteristics need to be considered.

- **Task 4: Model Evaluation and Result Reporting.** Due to challenges such as weak association and multicollinearity, there is no super model that can assess all the variables at the same time. Multiple models need to be built and evaluated interactively and progressively. The results need to be effectively conveyed to users so that they can conduct further model building and report the findings.

To effectively support the above tasks, I argue that my visual analytics system should have the following features: **(1) Integration:** The system should support the general workflow of risk factor analysis and carry out all the aforementioned tasks in

a seamless pipeline. **(2) Effective dimension reduction:** Intuitive visualization and flexible interactions should be provided so that users can efficiently examine a variety of indicators for a large number of variables to support the dimension reduction task. **(3) Transparent relationship analysis:** The reason a variable is included/excluded from a model should be transparent to users by explicitly providing the confounding and correlation information. **(4) Interactive model building:** The system should allow users to interactively build stable multivariate logit models based on dimension reduction, relation analysis, and domain knowledge. **(5) Complete and accurate results:** The system should effectively facilitate users in finding all potential risk factors from a high-dimensional dataset and contain as few false-positives as possible. **(6) Informative reporting:** The system should provide not only the names of risk factors, but also details and meaningful information on confounders and risk factors. **(7) Intuitive visual interface:** Targeting domain experts such as epidemiologists, the visual encoding should carry clear statistical meaning to users. Easy-to-use interactions should be provided to help the domain experts conduct the tasks.

3.4 Dimension Reduction

The entire visual analytics process consists of Dimension Reduction, Relationship Analysis, and Model Evaluation as in the Figure 9 indicated.

In this Dimension Reduction step, users interactively select interesting explanatory variables based on univariate indicators.

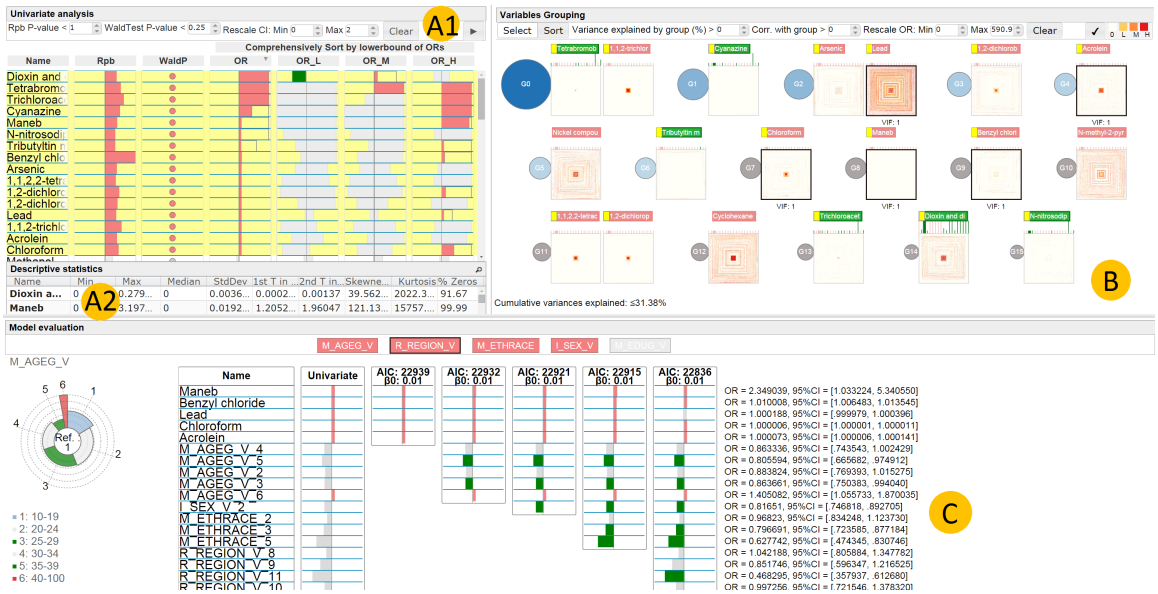

Figure 9: The interface of visual analytics for high-dimensional logistic model building for outcome of the limb reduction defect. (A1) The univariate analysis view. (A2) The descriptive statistics information panel. (B) The variable grouping view. (C) The model evaluation and comparison view.

3.4.1 Statistical procedures

Our prototype provides the following indicators frequently used in epidemiological analysis. They describe different aspects of the relationship between an explanatory variable and the birth defect variable.

(1) Point Biserial Correlations (rpb): See Chapter 2.3. The same statistical significance level of 0.05 is used. **(2) Wald test p-value (WaldP):** A WaldP [46] comes from a univariate logit model that contains a chemical exposure variable and a birth defect outcome. It is used for testing the statistical significance of the model. If the p-value is less than or equal to a chosen significance level, the chemical variable is doing much to help explain the birth defect outcome. Generally, the significance level is set as 0.05. **(3) Crude Odds Ratios (OR) and their confidence intervals:**

OR is a quantitative measure of the association between a binary explanatory variable X and a binary outcome variable Y (see chapter 2.1.4). ORs for chemical variables are calculated using the exponential function of the regression coefficient.

(4) Crude ORs based on categorized chemical exposure variables and their confidence intervals: To explore an association between an exposure and a birth defect beyond a yes/no (dichotomous) exposure, but not assuming a linear association with a continuous exposure (as in the point-biserial correlation), epidemiologists sometimes categorize the exposure using quartiles or some other quantile based on the reference group. Because a large proportion of the estimated exposures in this dataset were zero, we categorized non-zero exposure values as "low", "medium", or "high" where the number of observations in each group were approximately equal. We mark the odds ratios as OR_L, OR_M, and OR_H for the "low", "medium", and "high" groups, respectively.

3.4.2 Visualization and interactions

Users need to examine the individual indicators as well as the consistency among different indicators (e.g., all positive or negative) for a large number of variables. They also need to select variables of interest according to multiple indicators. Our system supports those tasks using the Univariate Analysis View (UAV) (Figures 9 (A1) and 10). UAV allows users to effectively examine the varying indicators for a large number of variables and select variables of interest flexibly and efficiently.

Since the indicators are measured in different ways, their visual representations are different. To help users judge them intuitively, a consistent color design is used

throughout the system: red indicates a statistically significant risk factor, green indicates a statistically significant variable that has been proven to be false-positive or protective, and gray means that the variable is not statistically significant. A variable may turn green from red as the analysis goes further.

As shown in Figure 10, the indicators are displayed in the UAV in a table-like view. Each row represents a variable whose name is displayed on the leftmost cell; each column represents an indicator. With the juxtaposition design [37], it is convenient for users to compare indicators/variables. Inspired by the rank-by-feature framework [82], I allow users to sort the variables according to one or more indicators and then select top ranked variables for further analysis. Following Table Lens [80], I make the inspection intuitive and effective by using visual attributes to represent indicator values.. Horizontal bars encode the rpb values. Zero values, the center of the rpb range, are placed at the center of the cell and marked by a short vertical line. Positive/negative rpb values are represented by a bar on the right/left of the vertical line; the length of the bar represents the absolute rpb value. If the p-value is greater than 0.05, the bar is colored gray. Otherwise it is red/green to indicate statistically significant positive/negative correlation. Significant/non-significant WaldP values are represented by a red/gray dot.

For OR, OR_L, OR_M, and OR_H, a horizontal axis is used in the cell and 1 is marked by a vertical line. If 1 is between the higher and lower bounds, the association is non-significant and I color the rectangle between the lower and higher bounds gray. It is desired to display a larger red/green portion in a cell of a more risky/protective variable. Therefore, I color the rectangle between the lower bound and the vertical

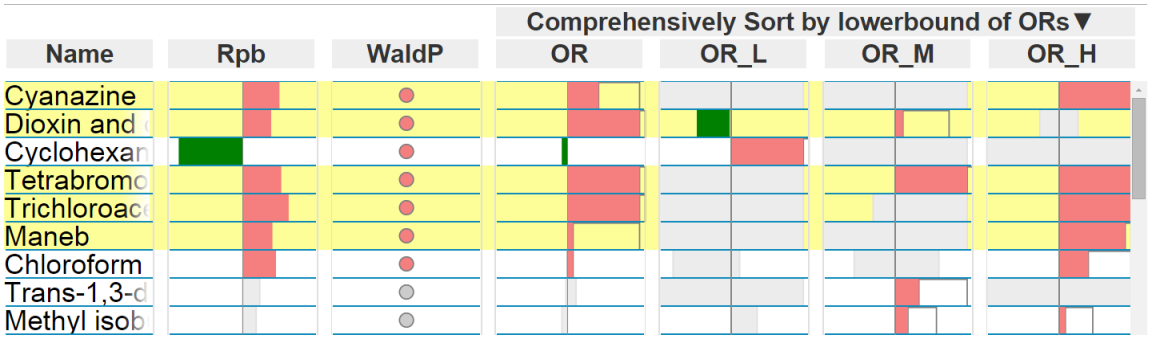

Figure 10: Indicators displayed in the Univariate Analysis View. The variables selected are highlighted in yellow.

line marking 1 red if the lower bound is higher than 1. I color the rectangle between the higher bound and the vertical line green if the higher bound is smaller than 1 (see Figure 10). This encoding makes it possible to rescale the cells because even if the confidence intervals extend beyond the cells, the risk/protective factors are still indicated by cell color (see Figure 10).

This view provides a set of interactions for dimension reduction. Users can filter out variables whose WaldP or rpb p-values are higher than a threshold. They can also sort the variables based on any of the indicators. An interesting interaction called comprehensive sorting is provided. It sorts the variables by the maximum of the lower bounds of OR, OR_L, OR_M, and OR_H. This is a useful interaction since a variable may have a strong association as long as any of the indicators are significant. Users can interactively click a variable to select/unselect it or use shift + click to bulk select. Basic descriptive statistics of the selected variables are displayed in the bottom of the UVA (see Figure 9 (A2)).

Users can send the selected variables to the next step by clicking a button. Since they can be unaware of risky variables correlated to a selected variable, the system

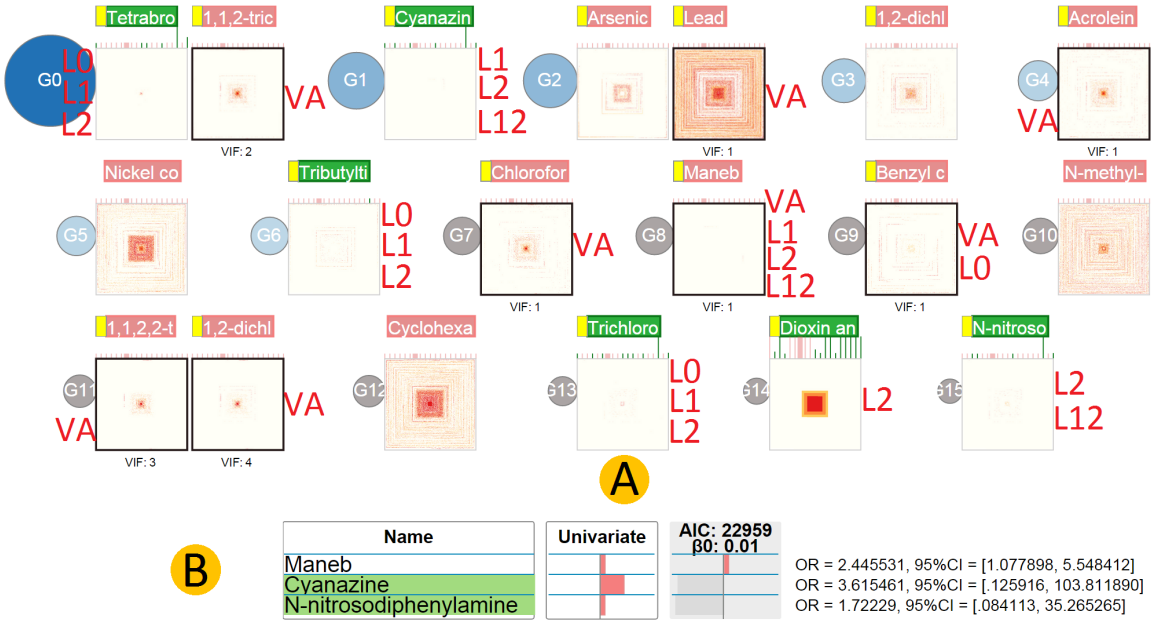

Figure 11: A. The variable groups view. L0, L1, L12, and VA indicate risk factors identified by Forward Stepwise selection, LASSO, Elastic Net, and my visual analytics approach, respectively. B. The result of a model with variables selected by L12. The gray in the rightmost column indicates that the model is not statistically significant. The green background behind the chemical names in the first column indicates the presence of confounding.

will automatically examine if such variables exist and add them to the next step. In particular, confounding tests (see Section 3.5) are conducted for each selected variable. All unselected variables which either confound or are confounded by the selected variable are included into the selection. Real-time interactions are possible because confounding tests for all chemical variables are conducted during pre-processing and the results are stored in a matrix. The system only needs to find the correct matrix location during the interactive selection. Variables selected in the UAV view by a user are marked in the next view by a small yellow block in front of the labels.

3.5 Relationship Analysis

In this step, users inspect the relationships among the variables from step 1 and interactively select variables to build logit models. Information such as correlations among the variables, variable stability, and confounding are automatically analyzed and visually presented to users.

3.5.1 Statistical procedures

Factor Analysis is often conducted to find intercorrelated variable groups, variances explained by each group, and the contribution of each variable in the group. This information can guide users to select variables that contribute more to the variability of the dataset thereby avoiding superfluous variables whose response patterns are caused by their association with an underlying latent variable. In addition, group information is helpful for identifying variables with weak associations whose significant associations can only be observed when other variables in the same group are excluded from a model (see Figure 12). Our system uses varimax rotation to impose a partition where each variable has a large correlation coefficient with the group it belongs to and small correlation coefficients with other groups [53]. Each group defines a "factor". The eigenvalue of a factor reflects its contribution to the total variance in the correlation coefficient matrix.

Confounding tests are conducted as follows: a bi-variate logit model is constructed for the response variable and each pair of explanatory variables. A explanatory variable having a change in OR greater than 10% compared to its univariate logit model OR in any bi-variate logit model is considered a confounder and the pairing variable

caused the confounding [44].

3.5.2 Visualization and interactions

In order to select variables for model building, users need to examine variable correlation, inspect Factor Analysis and confounding test results, and study data distributions of the variables. Data distribution is important for domain experts to understand the quality of data collection, inspect levels of exposures, and validate and hypothesize the correlation between variables. My system provides the Variable Groups View (VGV) for interactive variable selection which supports the above tasks (see Figure 11). Compared with the UAV, the VGV needs to provide many more details for a smaller number of variables.

Presenting the dataset is challenging since its large proportion of zeroes causes clutter in many visualization techniques. I selected the pixel-oriented technique [57] since it can present datasets with a large proportion of a single value without clutter. In addition, it scales to large, high-dimensional datasets and allows users to inspect variable distributions and dimension correlations at the same time. Other popular techniques such as scatterplot matrices [43] and Parallel Coordinates [48] do not have all of these characteristics.

In Figure 11, each square represents a variable. In a square, each data item is represented by a pixel whose color represents its value for the variable: white means zero; dark, medium, and light red represents high, medium, and low values of the exposure to that chemical (calculated with the same partitioning approach used in the univariate analysis), respectively. The squares can help users examine the distri-

butions of the variables. For example, Figure 11 shows that *lead* is widely distributed while *Cyanazine* is sparsely distributed. A data item has the same position in all the squares. Therefore, correlated variables have similar textures in their squares. We can observe that *1,1,2,2-tetrachloroethane* and *1,2-dichloropropane* are highly related from Figure 11. Users can sort the items by clicking a variable. The sorting will place items with high values for that variable in the center of the display. This allows the correlation between that variable and other variables to be clearly observed. In Figure 11, the records are sorted by *Dioxin and Dioxin-like Compounds*.

Variables are grouped and placed in the VGV based on the Factor Analysis results. To use the space efficiently without clutter, a grid layout is used and the variables are placed on the grid line by line. Variables of the same group are placed adjacent to each other. Different groups are bounded by circles. Groups with larger eigenvalues are considered more important and their circles are bigger and darker. Groups are sorted in descending order by eigenvalues. Within a group, a variable that is more correlated with the group is placed in front of variables that are less correlated.

Selection mode in the VGV allows users to add a variable into a logit model by clicking it. Dynamic tips for Variance Inflation Factor (VIF) [53] are provided during variable selection. It provides an estimation of the severity of multicollinearity in selected variables according to the correlation matrix. In addition, the cumulative variance of selected variables is also displayed at the bottom of the VGV during the variable selection process. In this way users can learn how much variability has been explained by the selected variables (see Figure 9 B).

For the confounding tests, analysts are interested in OR changes of 10% or more

in the N-1 (N is the number of variables in the VGV) bi-variate logit models in which a variable participates. Very large OR values are also interesting since they are an extreme case of confounding. I propose a compact grass view attached to each pixel-oriented display for visualizing the variable's OR values and changes in the N-1 bi-variate logit models (see Figure 11). N-1 grass blades (tiny vertical lines) grow on top of each square. Each grass blade represents the result of one logit model. If the OR change is 10% or more, the grass blade and the label are green to indicate that the variable is false-positive (a confounder). Otherwise the grass blade is red. The label of a variable is red if all its OR changes are less than 10% (it is not a confounder). The length of a grass blade represents the OR value. It is normalized among all the variables to enable comparison. Tall grass blades, which mean inflated ORs, can be easily spotted. Since extremely high OR values of the confounders may skew the OR distribution, I set the upper bound of the normalization to be the maximum of 20 and the largest OR.

When users hover a mouse over a grass blade, the blade of the paired variable is highlighted by an increased width. Users can examine the OR value or click the label of that variable to find out whether it confounds other variables (they will be highlighted). By comparing the textures of the squares, users can learn the correlation among the variables and thus get a deeper understanding of the confounding. Users can also examine the pairing variables one by one through a navigation widget triggered by clicking the grass.

Interactive visualization provides the flexibility of building models using different strategies. For example, users can start by adding non-confounders with distinct

pixel textures into the model. If the model is good (the background color of the result column in the Model Evaluation View is not gray), more variables whose group mates are not in the model can be included. In addition, the weak association between the outcome and a variable can be suppressed if its group mates are in the same model. Removing the group mates will allow the model to reveal the weak association (see Figure 12 for an example). Users can also use a full model with all non-confounders, where they only pick one variable from each group to build the model. Later, they can replace any variables with their group mates to check the effect of the group mates. At any time during interactive model building, if the model is not good (indicated by a gray column in the Model Evaluation View), users can remove the variables with unstable estimation from the model to improve its stability.

3.6 Model Evaluation

After users click a button in the VGV, a multivariate logit model will be built with the variables selected in it. The users can interactively examine the results of the model for refinement or for testing other variables. The categorical demographic characteristics can also be added into the model for effect change analysis.

3.6.1 Statistical procedures

The Newton-Raphson technique [97] is used to optimize logistic regression coefficient computations. Their results include ORs and confidence intervals for each variable. If a variable has a change in OR greater than 10% compared to its OR in the univariate logit model, it is considered a confounder. A non-confounder whose lower confidence interval bound is larger than 1 is considered a risk factor. If there

are one or more variables with large OR values (such as an OR larger than 20), the model is considered unstable and needs to be improved.

The Likelihood Ratio test [46] is used to measure the statistical significance of the model. The model is significant if the p-value is smaller than 0.05. Akaike's Information Criterion (AIC) is another measure to assess the goodness of fit [46]. The smaller the AIC, the better the model is.

To help users analyze categorical variables, a Chi-square independence test [46] is conducted. It is not interesting to study a categorical variable which has a non-significant association with the birth defect being studied. For each category in a variable, a contingency table is constructed using a reference category assigned by domain experts. ORs and confidence intervals of these groups are calculated to discover groups vulnerable to the birth defect.

3.6.2 Visualization and interactions

The Model Evaluation View (MEV) allows users to examine the results of a set of multivariate logit models. Besides the model consisting of all the variables selected from the VGV, users can interactively build more models with those variables and one or more categorical variables so that the effect change of the categorical variables can be evaluated. The results are presented in a table (see Figure 9 C). Following SAS [81], the first column of the table shows variable names and the other columns record the results of the models. The names of the confounders are highlighted in green (see Figure 11 B). The second column shows the ORs and the confidence intervals resulting from univariate logit models. The other columns show the ORs and the

confidence intervals resulting from the multivariate logit models. The ORs and the confidence intervals are displayed using the same visual encoding of the UAV. Their numeric values in the last model are displayed after the last column so users can read the results accurately. Once users have a stable model containing chemicals only, they can add categorical variables into the model to examine their effect change. As shown in Figure 9 C, the color of the categorical variable name indicates their significance in the Chi-square independence test (gray for non-significant variables and red for significant variables). Clicking the name of a categorical variable will trigger its Rose Plot (see Figure 9 C on the left). A Rose Plot visually presents the population sizes, the proportion of cases, and the significance of the ORs associated with each variable category. In particular, it consists of multiple sectors, each of which represents a population group defined by a category. The angle of a sector is the proportional to the number of mothers in this group. The radius is the ratio of cases to the number of mothers in this group. The red/green color indicates this population group has a significant OR with a confidence interval lower bound larger than 1/higher bound smaller than 1 (vulnerable/resistant to the birth defect). The reference group is colored blue and other groups are colored gray. The Rose Plot is compact and allows users to effectively compare populations and risks. It works well for variables with a small number of categories, which is the case in this application.

Users can gradually add the categorical variables into the model by double clicking them. The results are shown column by column. The effect change can be examined by comparing these columns. For example, as shown in the Figure 9 C, column 3, 4, 5, and 6 do not show any big differences in ORs or in the confidence intervals for

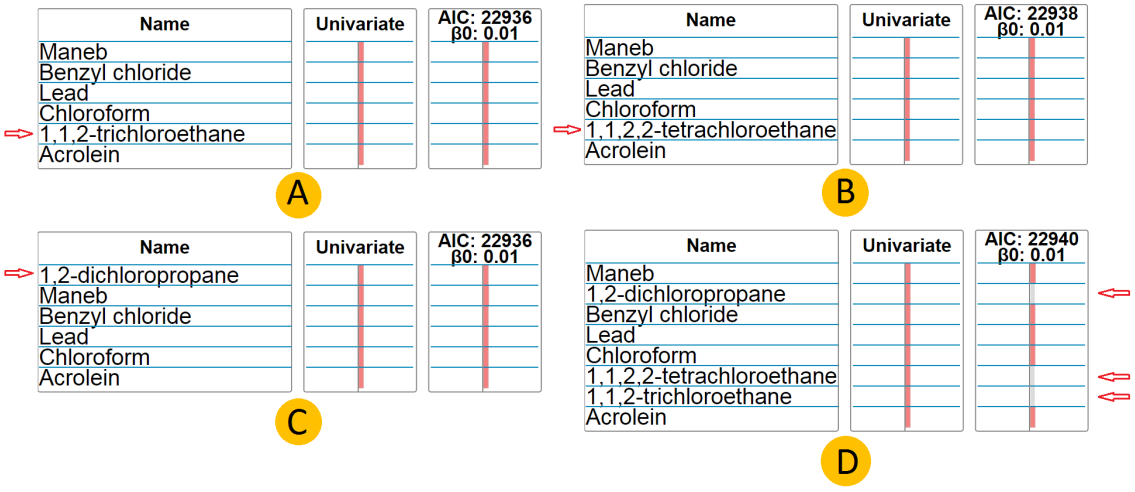

Figure 12: Weak associations. A. *1,1,2-trichloroethane* is included in a model with stable variables. It has a red bar indicating its statistical significance. B and C show the same pattern for *1,1,2,2-tetrachloroethane* and *1,2-dichloropropane*, respectively. D. The three of them are added into the same model. They all have gray bars which mean that they become non-significant.

the chemical *lead*. However, the effect for *lead* becomes non-significant in the last column. The column has an additional categorical variable *region* in the model. This tends to explain that the effect of *lead* varies by *region*.

Weak effects can be uncovered by interactive model building. For example, by looking at the data distribution in the VGV (Figure 11 A), I find three chemicals tightly correlated. They are *1,2-dichloropropane*, *1,1,2,2-tetrachloroethane*, and *1,1,2-trichloroethane*. A weak association between each of them and the birth defect can be identified by the model shown in Figure 12.

3.7 Use Cases

In this section, we first report a case study of identifying risk factors for Limb Reduction Defects (LRD) conducted by the authors. The results are compared with five automatic dimension reduction methods. Then we present a case study on the COIL

Challenge 2000 benchmark dataset [91] to characterize caravan policy holders. The study was pair analytics [8] by a senior Ph.D. student in Statistics and a visualization expert.

3.7.1 Identifying risk factors for Limb Reduction Defects

We conducted a case study for LRD using the environmental birth defect dataset introduced in Section 1. Sixteen variables were selected from the UAV (Figure 9 A1) with a 0.25 WaldP threshold ([17] and [46] suggested this threshold for dimensional reduction). The variables are sent to the VGV (Figure 9 B) together with three other variables correlated to them (the three variables were later proven to be non-risk factors). A robust model was quickly built. It identified five risk factors (the top five rows in Figure 9 C). Then, we further identified three risk factors having weak associations with LRD through interactive model building (Figure 12 illustrates this process).

To compare our approach with existing approaches, we fed the 16 variables selected from the UAV into 4 automatic approaches and compared their results with ours. They were Forward Stepwise selection [26], LASSO [89], Ridge Regression [18], and the Elastic Net [101]. The stats package [79] and the glmnet package [33] in R were used. The penalty parameter lambda was chosen based on cross-validation provided in the package.

The results favored our approach. First, only a small number of identified risk factors were consistent among the automatic approaches. Second, several risk factors identified by one or more automatic approaches were not identified in our case study.

We have proved that they are all confounders (see Figure 11 A for an example). Third, our approach identified several risk factors that were not identified by any automatic approaches. Among them, *lead* is a well-studied metal, so we conducted a literature search to find the ground truth. We found several articles [36, 67] suggesting that *lead* levels in hair and blood of mothers are related to LRD, which is consistent with our result that *lead* may be a risk factor for LRD.

3.7.2 Finding characteristics of caravan policy holders

The COIL 2000 Challenge benchmark dataset contains customer information from an insurance company [91]. It consists of 9,000 data items and 86 variables, including product usage and socio-demographic attributes. This dataset was selected because it represents many domain-specific problems: noisy data, correlated items, redundancy, high-dimensional variables, and weak associations between the explanatory and response variables. The task is to characterize caravan insurance policy holders and provide insights into why customers have the insurance.

The participants were Jean, a senior Ph.D. student in Statistics whose research area is variable selection and regression model building, and Joe, a senior Ph.D. student in visualization who is familiar with my visual analytics prototype. Both of them were not familiar with the dataset before the study. They did not know the semantic meaning of the attributes (the text inside [ ] following a variable name below) until the end of the study.

Jean and Joe explored the dataset side by side in front of a desktop where the dataset was loaded into the prototype. Joe briefly introduced the overall workflow to

Jean at the beginning of the study. Jean then took charge of the reasoning process and Joe helped her in manipulating the visual interface and explaining the visual encoding of the data displayed.

First, Jean filtered, sorted, and selected attributes through the UAV. She used 0.25 as the threshold for the rpb P-value and Wald test P-value. Thirty-two attributes were selected and sent to the VGV. Extra attributes were automatically sent to the VGV since they correlate to one or more variables she selected. From the textures of the squares in the VGV, Jean commented that she could tell that the data was redundant and correlated. She found several variable groups, such as a group consisting of MOSTYPE [customer subtype] and MOSHOOFD [customer main type], as well as a group consisting of MGEMOMV [avg. size household 1-6], and MFWEKIND [household with children]. She noticed that MKOOPKLA [purchasing power class] confounded the relationship between the caravan policy and MGEMOMV [avg. size household 1-6]. She also found correlations between attributes after sorting the textures. For example, MKOOPKLA [purchasing power class] had a negative correlation with the group of MOSTYPE [customer subtype] and MOSHOOFD [customer main type]. APERSAUT [number of car policies] and PPERSAUT [contribution car policies] had a positive correlation. After three iterations in the steps of model building, Jean obtained a good model with 21 variables. There were 13 non-significant and 8 significant attributes as the color coding indicated. Excluding 2 confounders indicated by a green background, Jean concluded that 6 significant attributes were likely to describe the characteristics of caravan policy holders. They were *PPERSAUT* [contribution car policies], *MAUT1* [1 car], *MBERMIDD* [middle management], *MO-*

PLHOOG [high level education], *PBRAND* [contribution fire policies], and *ALEVEN* [number of life insurances]. These attributes were commonly acknowledged in [91]. They were likely to describe a group of rich people with a more expensive car, a high level of education, and fire insurance due to the need to carry gas for cooking in the caravan.

3.8 Expert Feedback

The LRD case study and experiment results were shared with the epidemiologists in a written document. One epidemiologist commented that the models from the automatic approaches were behaving oddly. For example, one model had an OR of around 262 for *Tetrabromobisphenol A*. The epidemiologist said it was extremely unlikely to be valid. He suspected that the inconsistency among the results of the automatic approaches was more likely caused by the high correlation among the chemicals than caused by missing observations, since the data contains 60,613 cases.

The univariate analysis view was demonstrated to the epidemiologists. They were excited about the visualization and commented that it was intuitive and makes their tasks of comparing the indicators much easier. Since they are not experts in high-dimensional logit modeling, they suggested that I consult a statistician for feedback on the multivariate analysis part of the prototype.

I followed their advice and interviewed a statistician through Skype. He has a PhD degree in statistics and currently is a professor and active researcher in the field. He has conducted intensive research on high-dimensional logit modeling. The interview lasted one hour, before which he had read a written document illustrating

my approach. I showed him a live demo of the system in the interview. During the demo, I explained the statistical procedures and visualization techniques to him. He validated the statistics procedures I used and made the following comments:

"This is a cool and useful system."

"It allows statisticians to communicate with users much easier. It conveys the modeling process to users in a visible, intuitive, user-friendly way rather than using tedious word descriptions."

"It follows the high-dimensional logit modeling pipeline I use. It nicely integrates a variety of statistical procedures together for effective logit modeling."

"It allows users to compare results from different statistical procedures."

3.9 Discussion

In this chapter, I present a novel visual analytics approach to high-dimensional logit modeling for risk factor identification. It integrates a set of useful statistical techniques for dimension reduction and model building into a smooth analysis pipeline. It enhances the analysis process with intuitive visualizations and interactions that allow users to easily compare the results from varying statistical analyses, provides rich detail information such as variable correlations, data distributions, and detailed results of confounding tests and factor analysis, and allows users to effectively and efficiently conduct dimension reduction and model building iteratively.

Case studies have been conducted where the prototype has been used to find potential risk factors for limb reduction defects and characterize the caravan insurance policy holders. The results present the effectiveness and efficiency of my approach.

Positive feedback has also been received from two epidemiologists and a statistician.

CHAPTER 4: VISUAL EXPLORATION OF HIGH-DIMENSIONAL CATEGORICAL DATASETS

In this chapter, we introduce a new visual analytics approach to facilitating categorical dataset exploration. It leverages association rule mining to reduce and order dimensions for Parallel Sets visualization to make exploration analysis and association discoveries more efficient.

4.1 Introduction

4.1.1 Categorical Datasets

Categorical datasets refer to datasets whose dimensions are categorical variables. Categorical variables are also known as discrete or qualitative variables and are common in many areas. Categorical variables can be further divided into either *ordinal* or *nominal*. Ordinal variables have a clear order or ranking. For example, an educational level might be categorized as elementary school, high school, college, and graduate school. Nominal variables do not have any kind of natural order and as the term indicates, they are named. An example is types of real estate that can be categorized as office property, retail property, industrial property, and residential property.

In categorical data analysis, analysts are often interested in the associations between explanatory variables and outcome variables. For example, for the *Titanic* dataset [23], an analyst may ask the following questions: How many people in the category *child* on the *age* dimension were in the *yes* category on the *survived* dimen-

sion, the outcome variable? How many of them were in the *first* cabin? Did the rule of "women and children first" apply?

4.1.2 Parallel Sets Visualization

There are two approaches to visualizing categorical datasets. The first one is to convert categorical variables to numerical variables and utilize numerical data visualization methods for visualization. Since the categories are mapped to a limited number of numeric values, information overlapping and visual elements stacking problems usually exist [70]. In addition, the data conversion imposes false ordering to nominal variables, which may lead to false insights. The second approach is to use specific visualization techniques that work with categorical datasets. Only a few visualization techniques have been developed for categorical datasets. Among them, Parallel Sets, or ParSets [9, 61], is the most popular one. ParSets is a multivariate visualization technique designed for exploring possible sets and subsets that exist in the categorical data. Technically, it is a mix between Parallel Coordinates [48] and Mosaic Plots [34]. Figure 13 shows the *Titanic* dataset [23] in ParSets.

In Figure 13, four dimensions (*Survived*, *Sex*, *Class*, and *Age*) in the *Titanic* dataset are represented as horizontal axes. Each horizontal line represents all the categories across one dimension. The width of the line encodes the percent of observations in that category relative to the dimension. In this example of Figure 13, the top axis shows the distribution between the survived and the perished passengers. It indicates that the perished was the largest group of passengers. Starting with the first dimension, each of its categories is connected to each category of the dimensions below. These

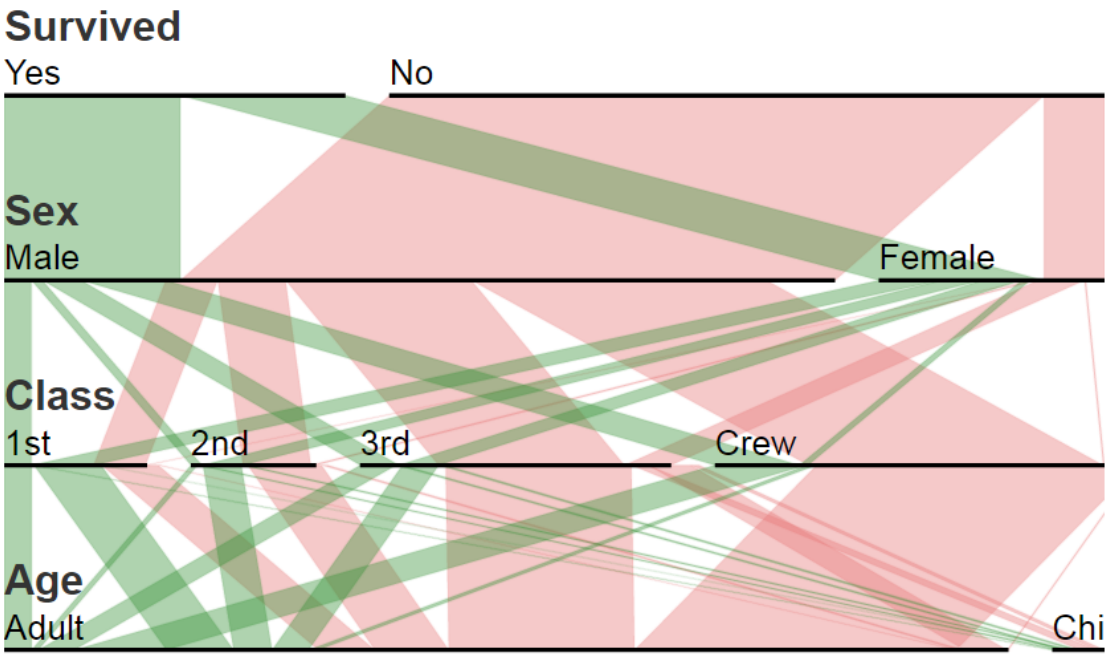

Figure 13: Parallel Sets for the Titanic dataset

connections form "ribbons" and continue across all dimensions showing how that categories are subdivided, how the combinations of categories are distributed, and how a particular subset (the women in the first cabin, for example) can be further subdivided (e.g., into those who were adults or children).

ParSets provides an interactive way to explore and analyze categorical data. The dimensions and categories can be rearranged to explore the relationships between different pairs of dimensions/categories. Typical interactions include dragging the dimensions and categories to order them.

4.2 Motivations

The key to successful visual analytics lies in the structure revealed within the data. However, a high-dimension data exploration often brings a crowed and disordered visual display (clutter) that obscures the pattern discovery and hinders the under-

standing of the data. The clutter problem is particularly serious in ParSets, which loses its effectiveness when there are more than several dimensions. Clutter occurs in ParSets when two or more ribbons are overlapped or pass through the same area. It is not a simple aesthetic problem. A severe clutter makes it hard to track, distinguish, and compare connections over more dimensions as continuity is lost. For example, in Figure 14, the ParSets for Mushroom dataset displays the dimensions and categories in alphabetical order. It is difficulty to find significant interesting patterns because of the severe clutter.

Many clutter reduction techniques have been proposed, such as dimension reduction approaches and dimension ordering approaches. Dimension reduction techniques include PCA [56], MDS [63], Factor Analysis [55]. They reduce clutter in visualizations through reducing the number of dimensions displayed. There are also geometric-based approaches [99], sampling approaches [27], and many other techniques [28].

Researches in the visualization community indicate that reordering dimensions in a visualization may reduce clutter and promote effective and efficient visual explorations. In the works presented in [95, 7, 78, 4], dimensions are reordered based on dimension similarities, clutter measures, or entropy to reduce clutter in visualizations such as parallel coordinates, ParSets, or pixel-oriented techniques.

These methods are very useful, but they are not targeted at help users identify the associations between dimensions/categories and dichotomous outcomes. For this specific task, new dimension reduction and ordering techniques need to be developed.

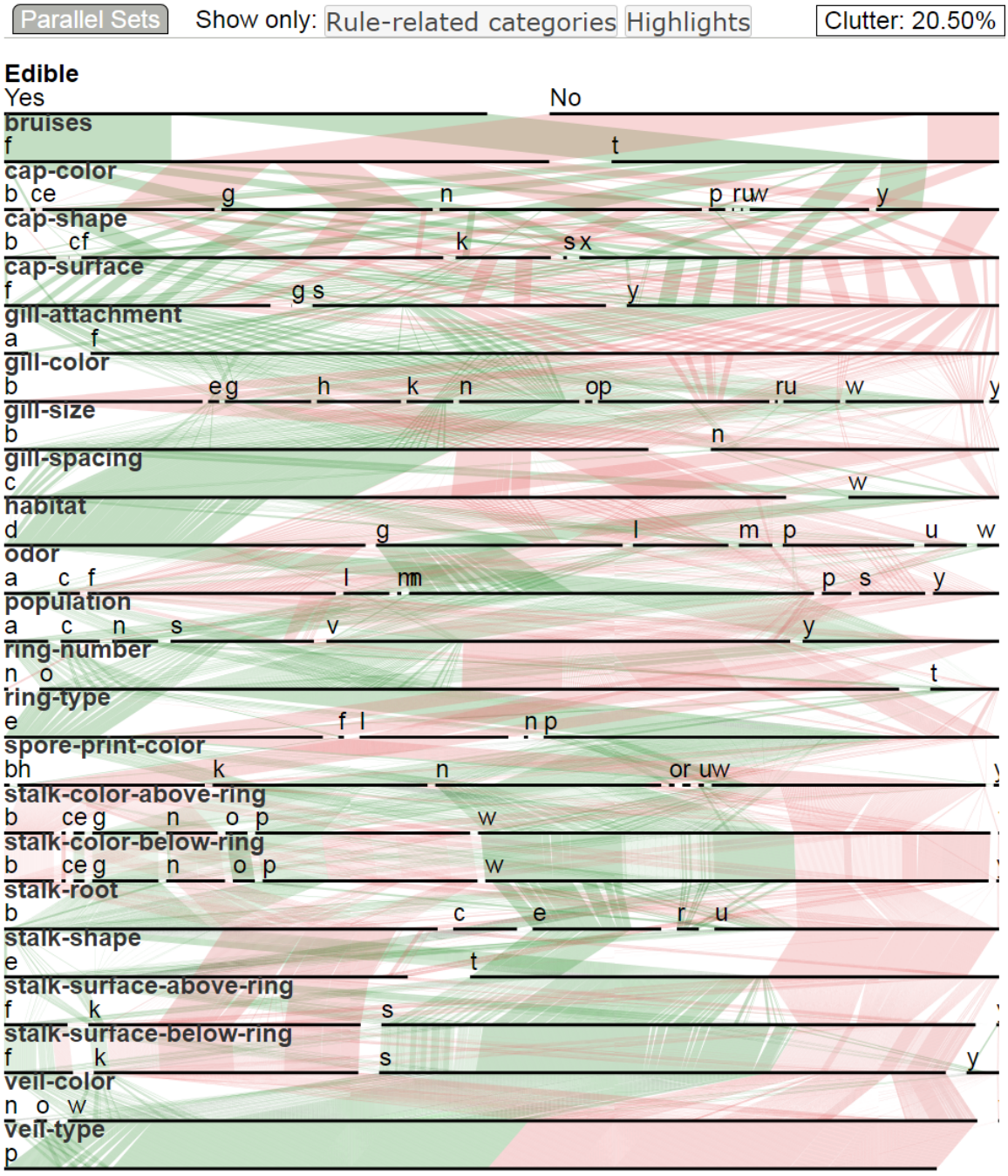

Figure 14: ParSets for Mushroom dataset displays the dimensions and categories in alphabetical order. The support range is from 18% to 34%.

4.3 Approaches and Contributions

In this chapter, we propose a new dimension reduction and ordering technique specifically designed for the task of helping users discover associations between di-

mensions/categories and dichotomous outcomes in ParSets. Instead of using dimension similarities or entropy, our approach orders and filters dimensions and categories according to their associations with the outcome, which are interpretable to human beings. In particular, we employ Class Association Rule (CAR) mining [71] to identify significant dimensions and categories related to the outcomes and hide insignificant variables from ParSets to reduce clutter. The dimensions and categories are sorted in ParSets based on their interestingness and associations in the association rules to leverage the visual exploration in ParSets using association rule mining results.

The main contributions of this chapter include:

- An interactive association rule view to help users identify dimensions and categories most associated with a dichotomous outcome. In particular, we use a table-like view to present rules distributed over the dimensions and categories. We color these rules based on the class (outcome) they imply to.

- A new interactive dimension and category reducing and ordering approach for ParSets. In particular, three approaches are proposed to order the dimensions based on category count, rule count, and closeness, respectively. In each approach, dimensions can be ordered differently according to different outcomes of interest. We also propose support-based ordering and confidence-based ordering to sort categories in ParSets.

- A set of case studies and experiments that illustrated the effectiveness of the proposed approaches in helping visual exploration and clutter reduction. A quantitative metric for clutter is proposed to measure the clutter reduction

effect. The proposed approach was compared with a recent entropy-based ordering approach [4] in the experiments.

4.4 Related Work

4.4.1 Categorical Data Statistics

Contingency table is commonly used when analyzing the relationship between categorical variables. For one categorical variable, we can summarize the data by counting the number of observation in each category. For two categorical variables, a rectangular table displaying counts for the two variables in which the rows represented one variable and the columns represent a second variable is called a two-way *contingency table*. Chi-Square test [64], Cramer's V [22], Phi Coefficient [21], Goodman & Kruskal's lambda [39], etc. can be calculated from the contingency table. They measure the relationship between the two variables.

The concept of PCA [53] has been applied to Correspondence Analysis (CA) and Multiple Correspondence Analysis (MCA) to project a categorical dataset to a two dimensional summarization display [45]. CA is performed on a contingency table associated with two categorical variables and decomposes the Chi-Square statistic of the table into orthogonal factors [2]. MCA is an extension of CA and allows one to analyze the relationship of several categorical variables in a two dimensional projection.

Categorical explanatory variables must be converted to dummy variables so that continuous regression modeling methods can be applied [2, 3]. Dummy coding uses only ones and zeros to convey all of the necessary information. The coded variables

uses one degree of freedom, so a variable having k categories will be coded into $k-1$ binary variables. Dummy coding apparently increases the number of dimensions. There is also some redundancy in the process of dummy coding [49].

4.4.2 Association Rule Mining

In the automated analysis field, association rule (AR) mining is often used for discovering co-occurrence relationships between categorical variables. Each rule is composed by two different sets of items, also known as *itemsets*, X and Y. The generic form of a rule is X $\Rightarrow$ Y, where X is called antecedent or left hand side and Y consequent or right hand side. For example, in the rule $\{onions,\ potatoes\} \Rightarrow \{burger\}$ found from a supermarket dataset, the left hand side is *onions* and *potatoes*, and the right hand side is *burger*. It indicates that if customers buy onions and potatoes, they are more likely to also buy a burger. There are several measures to evaluate the rule significance such as *support*, *confidence*, and *lift*. Support is a measure of how frequently the itemset appears in the dataset. Confidence is a measure of how often the rule has been found to be true. The confidence is the proportion of the transactions that contain X and Y together. It is defined as $conf(X \Rightarrow Y) = supp(X \cup Y)/supp(X)$. Lift is the ratio of the observed support to that expected if X and Y are independent. It is defined as $lift(X \Rightarrow Y) = supp(X \cup Y)/(supp(X) \times supp(Y))$.

In traditional AR, any item can appear as a condition or consequent of a rule. There is a special association rule called the Class Association Rule (CAR) [71]. It has a fixed item on the right side of the association rule as a consequent. It use the same generating and pruning algorithms as the traditional AR. The difference is

the algorithms only output rules having the specified right hand side. Our approach employs CAR since the label of the outcome of interest is the class of CAR. We are only interested in the related variables to that label.

AR mining has been applied to feature selection approaches [59, 19]. Ko et al. [59] apply it to word-list selection for classification in a high-dimensional document space. Backed by experiments, the authors demonstrate that AR mining outperforms information gain [74] and document frequency [65] in the classification task. To solve the problems of high dimensionality and small sample size in the neroimage field, Chaves et al. [19] propose an AR-based feature selection method for their computer aided disease diagnosis system. There are several differences between these works and our approach. First, the goal of our approach is to discover associations for explanation rather than classification, even though it can also be used for classification. Second, we focus on using visualization to understand the change of association in different circumstances, e.g. the combination of different attributes.

4.4.3 Categorical Data Visualization

In the visualization community, Friendly surveyed many visualizations for categorical data [35]. They include fourfold displays, mosaic plots, unordered histograms, and pie charts. According to [30], there are two categories of categorical data visualizations. They are QuantViz and CatViz. QuantViz needs a quantification process before the categorical variables are visualized [51, 50]. As discussed earlier, this category of techniques suffer from the clutter problem and misleading insights. CatViz directly employs visualization to frequencies or categories of categorical data. Exam-

ples include the Mosaic Display [34], CatTree [60], and Contignecy Wheel ++ [5]. ParSets [9] studied in this dissertation belongs to CatViz.

The Mosaic Display [34] visualizes multi-dimensional contingency tables using the n-way mosaic. CatTree [60] uses treemaps to visualize the hierarchies of the categorical dataset based on the frequency of a category. Parallel Sets [9] present frequencies as stripes between axes in a layout similar to Parallel Coordinates [48]. Contingency Wheel [6] and Contingency Wheel ++ [5] are directly based on tables of category frequencies (contingency tables). Based on the result of MCA, Broeksema et al. propose an interactive voronoi diagram to represent the clusters of similar observations and related attributes [16].

A significant drawback of the CatViz techniques is that they are weak in terms of high-dimensional categorical data exploratory analysis. When the dimensionality of the data increases, the display often becomes cluttered. Clutter reduction is needed in a frequency/category-based data visualization for effective insight discovery. In addition, contingency table-based visualizations, such as Contingency Wheel, is limited in exploring how the data frequencies are related across different dimensions.

4.4.4 General Clutter Reduction Techniques

There has been several discussion regarding the general clutter reduction techniques. Wegman and Luo [92] uses hue and opacity to encode clustering of lines for Parallel Coordinates, so that overplotted regions areas of high density become noticeable while outliers can be more readily distinguished. Ellis and Dix [27] argue that when there is too much data to be visualized, a random sampling might reduce

clutter while preserving the main trends of the data. However, it was accepted that the sample may not be an exact representation of the original dataset.

It has been shown that dimension reordering in a parallel coordinates reduces clutter [7, 95, 78]. Clutter is reduced when the data is well clustered between dimensions [78]. In most visualization tools dimensions are initially displayed in no particular order, usually in the same order they appeared in the input dataset. Manually rearranging the dimensions can lead to the discovery of relationships between neighboring dimensions but this can be a time-consuming task. The calculation of the number of possible orderings is a NP-complete problem and thus heuristic ordering methods have been proposed [7, 95].

For example, Yang et al. [95] present a similarity-based dimension hierarchies to interactively reorder hierarchical dimensions for Parallel Coordinates. Ankerst et al. [7] heuristically cluster data dimensions according to their similarity, then rearrange these dimensions. Alsakran et al. [4] consider the variation or diversity of dimension and use entropy to order axes in ParSets. Few of them are applied to categorical data visualization for finding associations between variables and outcomes.

To evaluate the effect of clutter reduction, it is necessary to measure how cluttered a visualization is. For visualizations where parallel axes are used, such as Parallel Coordinates and ParSets, the measurement is typically done by measuring the clutters between each pair of neighboring dimensions and then averaging those values across the whole visualization.

Peng et al. [78] proposes an algorithm in which the measure of clutters between dimensions is defined as the ratio of outlier points to the total data points. Using

the Euclidean distance, a data point is considered an outlier if it has no neighboring data point within a given threshold distance. The optimized dimension order is then computed to minimize the proposed clutter measure. In this dissertation, we propose a new clutter measure tied to the analysis task of identifying explanatory variables associated with a dichotomous outcome variable.

4.5 Approach Overview

We propose a new visual analytics approach and develop a fully working prototype for it (see Figure 15 for the interface). In this approach, Class Association Rules mined from a high-dimensional categorical dataset are used to interactively order dimensions and categories in ParSets as well as filter insignificant dimensions and categories from ParSets. This approach supports the following interactive exploration pipeline:

- First, Class Association Rule mining is applied on a categorical dataset with user-specified support, confidence, and lift thresholds. A set of associated rules are generated, which will be used in the visual exploration followed.

- Second, the users interactively examine the associated rules in a table-like visualization and select rules of interest. They can also prune undesired rules using support threshold filters.

- Third, the users select dimension and category ordering criteria based on their analysis tasks. The dimensions and categories in ParSets are then ordered based on the rules and ordering criteria selected. Only dimensions in the ARTable are displayed in ParSets.

- Fourth, the users hide insignificant categories from ParSets to reduce clutter. The significance of the categories are defined by the association rules.
- Fifth, the users discover interesting associations/patterns from ParSets and highlight them.
- Sixth, the users bring back hidden dimensions and categories and examine the discoveries in the context of the whole dataset.
- Last, the users search and filter the original dataset using Rawdata View to inspect or confirm with the most specific data records.

Users can go back to any previous steps in an iterative visual exploration. In the following sections, we illustrate the components in this visual analytics pipeline one by one. Before that, we first introduce a clutter measure in ParSets.

4.6 A Clutter Measure for ParSets

We propose a new clutter measure for ParSets. The basic idea is that we measure the clutters between each pair of neighboring dimensions, sum them up, and normalize the total clutters between 0 to 1 (100%). The worst scenario of clutter between neighboring dimensions is that categories of source dimension head-to-tail connect the categories of target dimension. Each connection (ribbon) then has the maximum distance between the two dimensions. In this case, the clutter of the ParSets becomes the most excessive and can be calculated as:

$$Clutter_{max} = \sum_{i=1}^{N} \sum_{j=1}^{M} D_{max}(src, tgt)$$

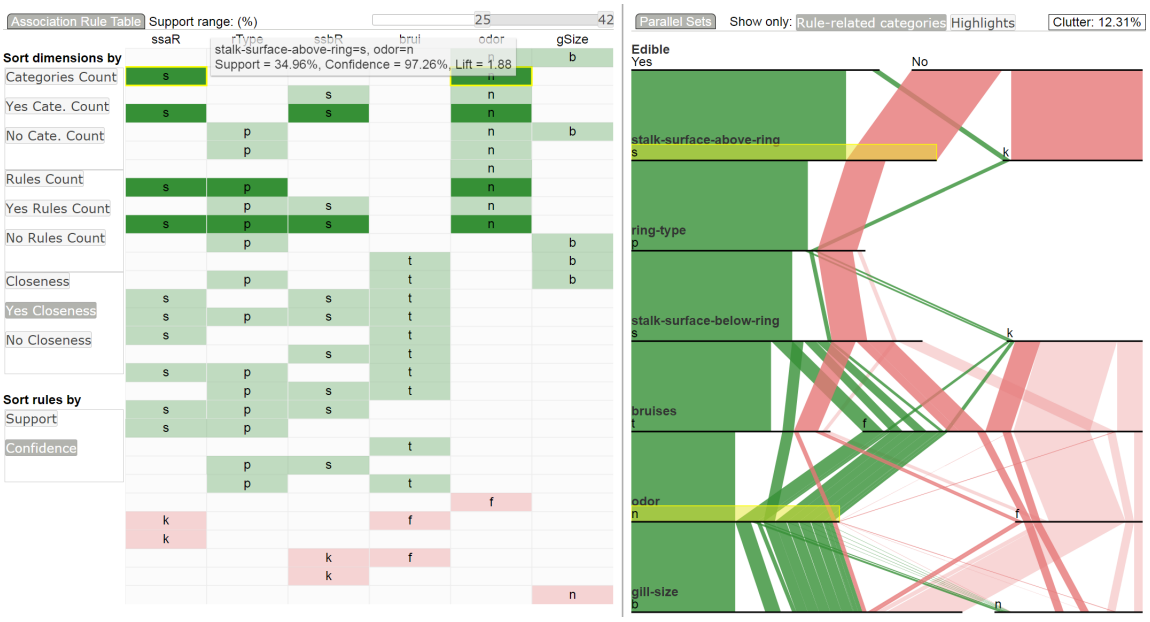

Figure 15: The interface of ARTable and Parallel Sets. In ARTable, the top rule $\{stalk\text{-}surface\text{-}above\text{-}ring = s,\ odor = n\}$ is clicked and highlighted in yellow border and dark green. Other rules sharing the same itemset are also highlighted in dark green. Meanwhile, the corresponding categories are highlighted in the ParSets as well as the ribbons that pass though the dimensions and categories that itemset contains.

The best case is that all ribbons between neighboring dimensions go straight from source dimension to target dimension without any obliques. In this case, the clutter of the ParSets can be calculated as:

$$Clutter_{min} = \sum_{i=1}^{N} \sum_{j=1}^{M} D_{min}(src, tgt)$$

The clutter in a display is calculated from:

$$Clutter = \sum_{i=1}^{N} \sum_{j=1}^{M} D(src, tgt)$$

, where N is the number of pairs of neighboring dimensions in ParSets, M is the number of ribbons between the neighboring dimensions, $D(src, tgt)$ is the Euclidean distance between the source and target category, $D_{min}(src, tgt)$ and $D_{max}(src, tgt)$

are constant given the ParSets width and between-dimension space, and $D_{min} \leq D(src, tgt) \leq D_{max}$. Then the normalized clutter in the ParSets is obtained using a min-max scaling which ranges from 0 to 1 (100%).

$$Clutter_{norm} = \frac{Clutter - Clutter_{min}}{Clutter_{max} - Clutter_{min}}$$

As shown in Figure 14, with the alphabetical order of dimensions and categories, the $Clutter_{norm}$ reaches 20.50%.

4.7 Association Rule Generation

A high-dimensional categorical dataset might result in a large number of association rules. Some of them are not of interest and need to be pruned either in or after the mining process using criteria such as *support*, *confidence*, or *lift*. Support is an important measure as described in chapter 4.4 because a rule that has a very low support may occur simply by chance. Thus, low-support rules have little contribution to the outcome occurrence. Different categorical datasets may need a different support threshold. Confidence measures the reliability of the inference made by a rule. It also provides a rule pruning method.

Common association rule mining can be broken down into two subtasks: *Frequent Itemset Generation* and *Rule Generation*. Frequent Itemset Generation aims at finding all the itemsets that satisfy the support threshold. These itemsets are called frequent itemsets. Rule Generation is to extract all the high-confidence rules from the frequent itemsets found in the previous step. We use the Apriori algorithm implementation by [11]. Because we are only interested in Class Association Rule [69],

we set the right hand side of association rules being generated to have one of the dichotomous outcomes at each mining.

For example, in the Mushroom dataset, we set the generating rules to have an appearance like $\{X\} \Rightarrow \{Y\}$, where the left hand side X is a subset of every possible attribute-value pairs except the outcome variable *Edible*, and the right hand side Y is a subset of the outcome variable *Edible* attribute-value pairs $\{Edible = Yes, Edible = No\}$. We have two example rules A: $\{bruises = t, gill-size = b\} \Rightarrow \{Edible = Yes\}$ and B: $\{bruises = f, gill-size = n\} \Rightarrow \{Edible = No\}$ (rule measures are omitted). Although rule A and B have the same attribute names, they lead to distinct rule classes because of the different attribute values.

4.8 Association Rule Table

We develop a table-like view, named ARTable, to visualize the association rules. ARTable allows users to examine the relationships between the association rules and the dimeisons and categories. It also serves as a control for manually ordering dimensions and categories in ParSets - the dimensions and categories in ARTable have the same order as in ParSets; it is easy to manually change the orders of dimensions and categories in this table-like visualization.

Figure 16 shows an ARTable of the Mushroom dataset. Each row represents an association rule and each column represents a variable. A cell is colored if the variable appears in the left hand side of the association rule (no matter what categories it has in the rule). The table is horizontally divided into two parts: the top part displays *edible* rules and the bottom part displays *poisonous* rules. To help users differentiate

the dichotomous outcomes intuitively, a consistent color design is used throughout the system: red indicates a *risk*, *negative*, or *bad* outcome, green indicates a *safe*, *positive*, or *good* outcome. For example, in Figure 16, red means rules for poisonous mushrooms and green means rules for edible mushrooms. The name of a category is displayed in a colored cell if its height is enough for showing the label.

With the juxtaposition design [37], it is convenient for users to compare dimensions and rules. For example in Figure 16, we can see that *odor* does not have any categories together with *brui* (bruises for full name) generating edible rules. Also, rules sharing itemsets can be easily found.

Although the rule generation process has already applied rule prune criteria (see chapter 4.7), a large number of rules can still be generated and sent to the ARTable. Some of them may be more interesting to users than other based on the analysis tasks. Therefore, we allow users to interactively filter the rules. In particular, a set of filters are displayed on the top of the ARTable. Users can use the filters to select rules applied in the dimension reduction and ordering process. Only selected rules and their relevant dimensions and categories are displayed in the ARTable. Support criterion is provided with range filters. We provide this criterion since different support thresholds may be preferred in different datasets. For example, in an income dataset, the high-income people have a smaller percentage and a lower support threshold needs to be used for a better coverage of their characteristics. Two support filters are provided so that users can set different support thresholds for different outcomes. We did not provide a color coding for the measures of support or confidence in an ARTable because all the filtered rules will participate in dimension ordering and

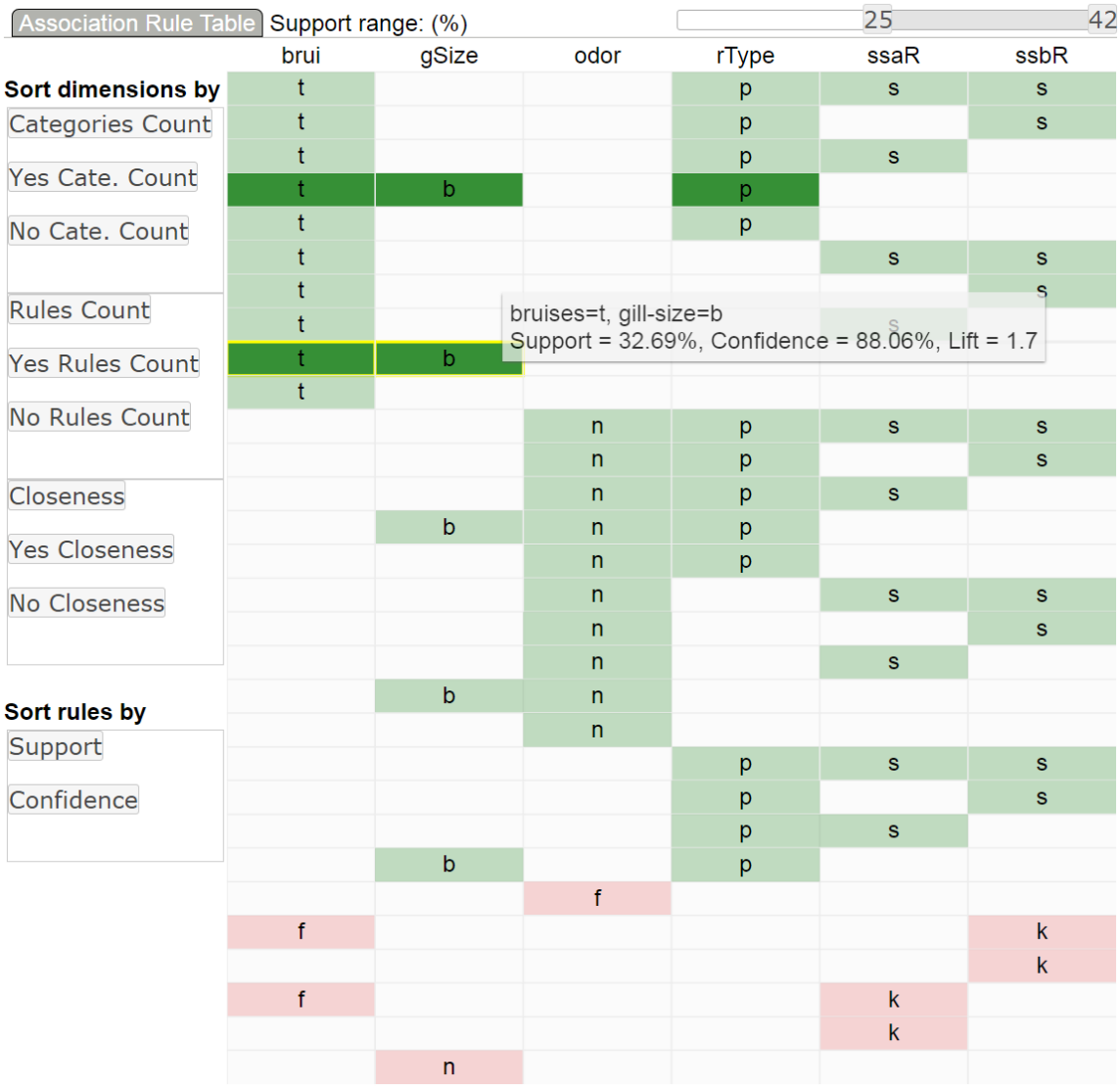

Figure 16: The ARTable view for the Mushroom dataset. Each row represents a rule and each column represents a categorical variable that is extracted from the association rules. The table cell is colored green or red according to the class (outcome of edible or poisonous) of the rule.

dimeison reduction.

A set of interactions are provided in an ARTable. For example, users can click a rule to highlight other rules sharing itemsets. This helps examine the rule relationships and identify rules that share the same itemset with a little extras. For example, in Figure 16, the clicked rule has the itemset of $\{bruises = t,\ gill\text{-}size = b\}$ and it is

highlighted in a yellow border. Two other rules are highlighted in a darker green fill, which indicates they are containing the same itemsets as the clicked one. Meanwhile, the highlighted itemsets cause a highlighting for the same attribute-values in the ParSets (see chapter 4.12).

4.9 Dimension Ordering

In this section, we propose three dimension ordering methods, taking advantage of the summarization ability of association rules. The first method orders dimensions based on their associations as revealed in the association rules. The second and the third methods rank dimensions based on two significance measures derived from the association rules. We introduce the three methods in the following sections, respectively.

4.9.1 Dimension Ordering by Associations

In this approach, dimensions with categories constructing the same frequent itemsets in the association rules are placed adjacent to each other, in the hope that there are wide ribbons (each band representing obervations in the same frequent itemset associated with a dichotomous outcome) connecting categories in adjacent dimensions in ParSets. In this way, the clutter can be reduced, frequent itemsets can be visible, and the interactions among associated dimensions and categories with regard to the outcome can be observed.

Following Hierarchical Dimension Ordering (HDR) [95], we construct a distance matrix for the dimensions, conduct a hierarchical clustering on the dimensions based on the distance matrix to group dimensions that are close to each other to the same

cluster, and sort the dimensions based on a breath-first traversal on the hierarchy.

In HDR, the distance matrix is built based on correlations among the dimensions. A different approach is used in our approach. Our idea is that having categories constructing the same frequent itemset in an association rule is the most important association between two dimensions. A pair of dimensions satisfies this criterion should be placed as close as possible. The distance matrix building process is illustrated below.

Given a set of N dimensions found from the ARTable, an N X N distance matrix is constructed in the following way: a. Select an outcome of interest. This step can be ignored if a user is interested in both outcomes; b. In the ARTable, if a cell has the color encoding the outcome of interest, then we put 1, otherwise we put 0 into this cell. If both outcomes are of interest, we put 1 in colored cells and put 0 in the white cells; c. Each column of the ARTable now forms a binary vector for the categorical variable. d. The distance between two vectors are calculated using Jaccard distance [66]. The binary matrix building process is illustrated in Figure 17 assuming the interest of outcome is *edible* mushroom.

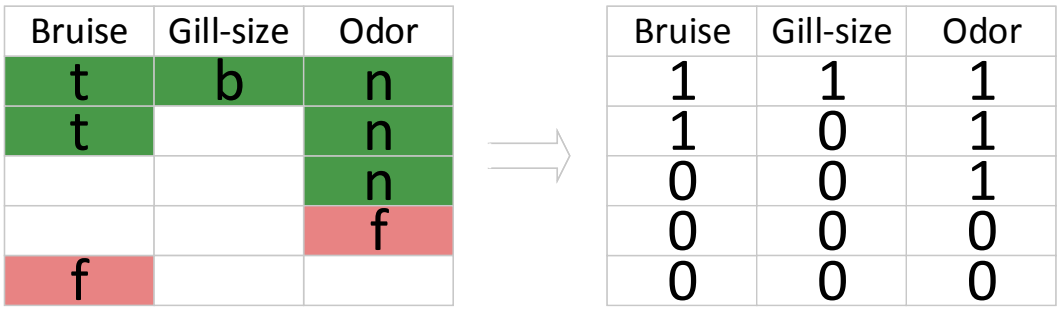

Figure 17: Binary matrix construction from ARTable for use in hierarchical clustering

The hierarchical clustering algorithm is very generic and conducted as follows:

1. Assign each dimension to a cluster containing only this dimension. So if we have N items, we now have N clusters, each containing just one item. Let the distances between the clusters equal to the distances between the dimensions they contain.

2. Find the closest two clusters and merge them into a single cluster. So now we have one less cluster.

3. Compute distances between the new cluster and each of the old clusters. The calculation is discussed later.

4. Repeat steps 2 and 3 until all the dimensions are in a single cluster of size N.

Step 3 can be done in different ways to determine how close two clusters are [54]:

a Single linkage: consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. It's strength is the ability to find irregular-shape clusters. The limitation is the sensitivity to noise and outliers.

b Complete linkage: consider the distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster. This approach is less sensitive to outliers than the single linkage approach.

c Average linkage: consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any

member of the other cluster. It is robust to outliers and tends to break large clusters.

In the resulting hierarchy, the leaf nodes are the dimensions. A breath-first traversal is conducted and the dimensions are ordered in the same order they appear in the traversal. This is a heuristic approach to placing close dimensions adjacent to each other [95].

4.9.2 Dimension Ordering by Category Count

The method considers placing dimensions generating wide ribbons on the top. It counts the number of categories of each dimension in the resulting association rules and sorts the dimensions by this *category count* in ascending order. Users can select to count the number of categories that only occur in the *yes*, *no*, or both class of the association rules. Since the outcome variable is always displayed as the first dimension (on the top of the display) in ParSets in our approach, hopefully this method can result in wide ribbons connecting the outcome(s) of interest and categories in dimensions ranked top in the display. For example, in Figure 18 A, dimension *Sex* having two categories, *Class* having four categories, and *Age* having two categories are adjacent in a default order that appears in the original dataset. The crowd ribbons have a clutter metric of 21.60%. While as shown in Figure 18 B, we sort these dimensions based on the number of categories, the clutter gets reduced to 17.33% and interpretation gets easier.

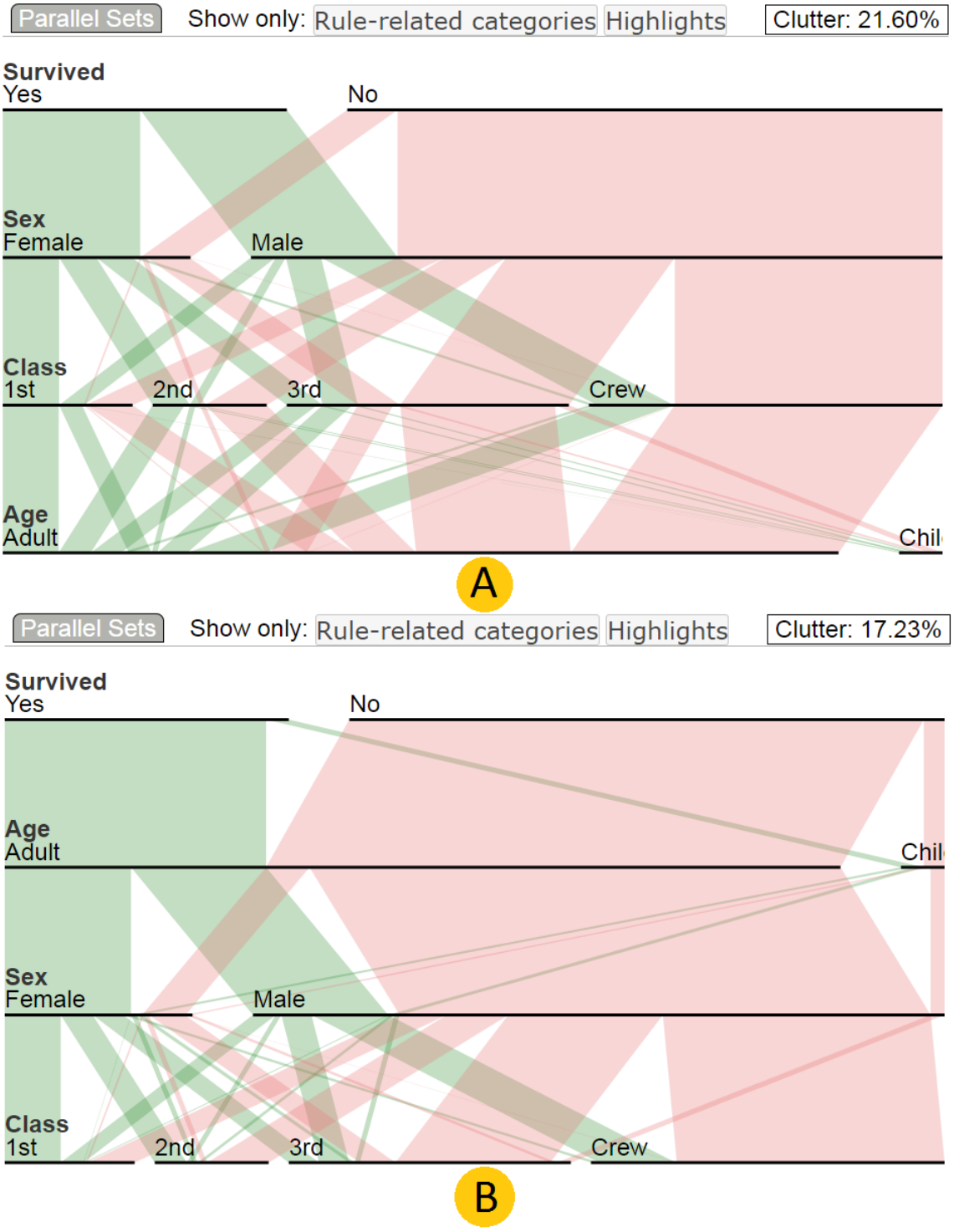

Figure 18: ParSets for the Titanic dataset - ordering with Category Count. A: a default order that appears in the original dataset. The clutter is 21.60%. B: ordered ParSets with Category Count and the clutter gets reduced to 17.33%

4.9.3 Dimension Ordering by Rule Count

This method orders dimensions based on the number of association rules they are involved. The assumption is that if a variable presents in much more association rules than other variables, it is more closely related to the outcomes and thus should be displayed closer to the outcomes than other variables. Similar to previous methods, users can count the rules in *yes*, *no*, or both classes of association rule.

4.10 Category Ordering

Besides dimension ordering, we also propose a category ordering technique to keep ribbons flowing vertically and reduce clutter in ParSets. Assuming that the outcome variable has two categories *yes* and *no*, we place them on the left and right side of the axis, respectively. Our idea is to place categories related to *yes* on the left of their axes and categories related to *no* on the right of their axes, so that there will be less ribbons running from left to right or from right to left to reduce clutter. Based on this idea, we propose the following algorithm to sort categories within each dimension:

a Sort all categories only appearing in *yes* rules by a selected rule metric (support/confidence) in descending order.

b For categories that exist in both *yes* rules and *no* rules and whose maximums of the metric in the *yes* rules is larger than or equal to that in the *no* rules, sort them by the metric in descending order. Append them after the last category in the previous step.

c For categories not in any rules and who have less *yes* observations than *no*

observations, sort them by the number of *yes* observations in descending order. Append them after the last category in the previous step.

d For categories not in any rules and who have more *no* observations than *yes* observations, sort them by the number of *no* observations in ascending order. Append them after the last category in the previous step.

e For categories that exist in both *yes* rules and *no* rules and whose maximums of the metric in the *no* rules is larger than that in the *yes* rules, sort them by the metric in ascending order. Append them after the last category in the previous step.

f Sort all categories only appearing in rules with *no* on the right side by the metric in ascending order. Append them after the last category in the previous step.

The algorithm is implemented as shown in Algorithm 1.

Figure 19 shows a clutter-reduced ParSets using *"yes"* closeness to sort dimensions and rule confidence to sort categories.

4.11 RawData View

RawData view is provided to examine the finest level of detail. It displays all the dimensions and all the observations from the original dataset in a table. Figure 20 shows the RawData View for the Titanic dataset [23].

Sorting, paging, and filtering are provided in the RawData View. Users are able to sort the observations by their values in a dimension by clicking its column header.

Algorithm 1 Category sorting algorithm in ParSets

```
1: A ← {all dimensions with ordered categories in ARTable} // see Algorithm 2.
2: P ← {all dimensions in ParSets}
3: for each pd ∈ P do
4:    ad ← find dimension matching pd.name from A
5:    if ad is TRUE then
6:       pdc ← {all categories in pd}
7:       adc ← {all categories in ad}
8:       updc ← {x ∈ pdc : x ∉ adc}
9:       for c ∈ updc do
10:         summarize the count of data items for each class
11:         determin the group of class that cbelongs to
12:      end for
13:      Y_o ← find categories mathing group = yes from adc
14:      N_o ← find categories mathing group = no from adc
15:      Y_u ← find categories mathing group = yes from updc
16:      N_u ← find categories mathing group = no from updc
17:      Descending sort Y_o by the selected measure
18:      Ascending sort N_o by the selected measure
19:      Descending sort Y_u by the summarized count
20:      Ascending sort N_u by the summarized count
21:      concatenate the four groups in the order of Y_o, Y_u, N_u, N_o
22:   end if
23: end for
```

Algorithm 2 Category sorting algorithm in ARTable

```
1: R ← {all current ordered rules in theARTable}
2: D ← {all dimensions in R}
3: L ← {yes, no}
4: for each d ∈ D do
5:    C ← {all categories in d}
6:    for each c ∈ C do
7:       for each cls ∈ L do
8:          c.cls_max ← Max of the selected measure
9:       end for
10:      c.group ← the class which has the maximal selected measure
11:   end for
12:   Y ← find categories mathing group = yes from adc
13:   N ← find categories mathing group = no from adc
14:   Descending sort Y by the yes_max measure
15:   Ascending sort N by the no_max measure
16:   concatenate the two groups in the order of Y, N
17: end for
```

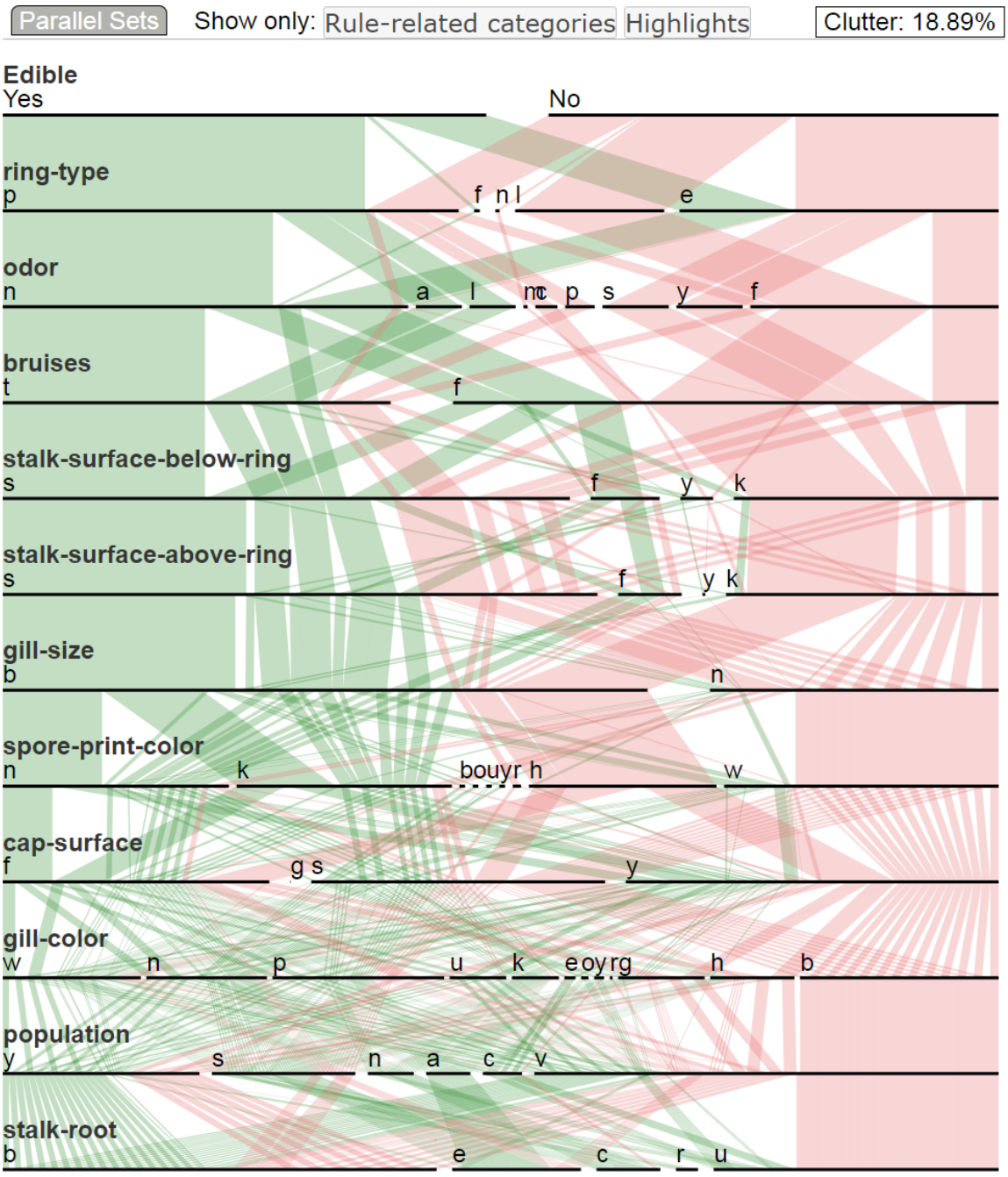

Figure 19: Clutter-reduced ParSets. The dimensions are sorted by *Yes Closeness* and the categories within a dimension are sorted by *Confidence* with the display of all categories. The support range is from 18% to 34%.

Paging is used for displaying a larger number of observations. Users can filter the observations with categories to examine the quality of an association rule or verify

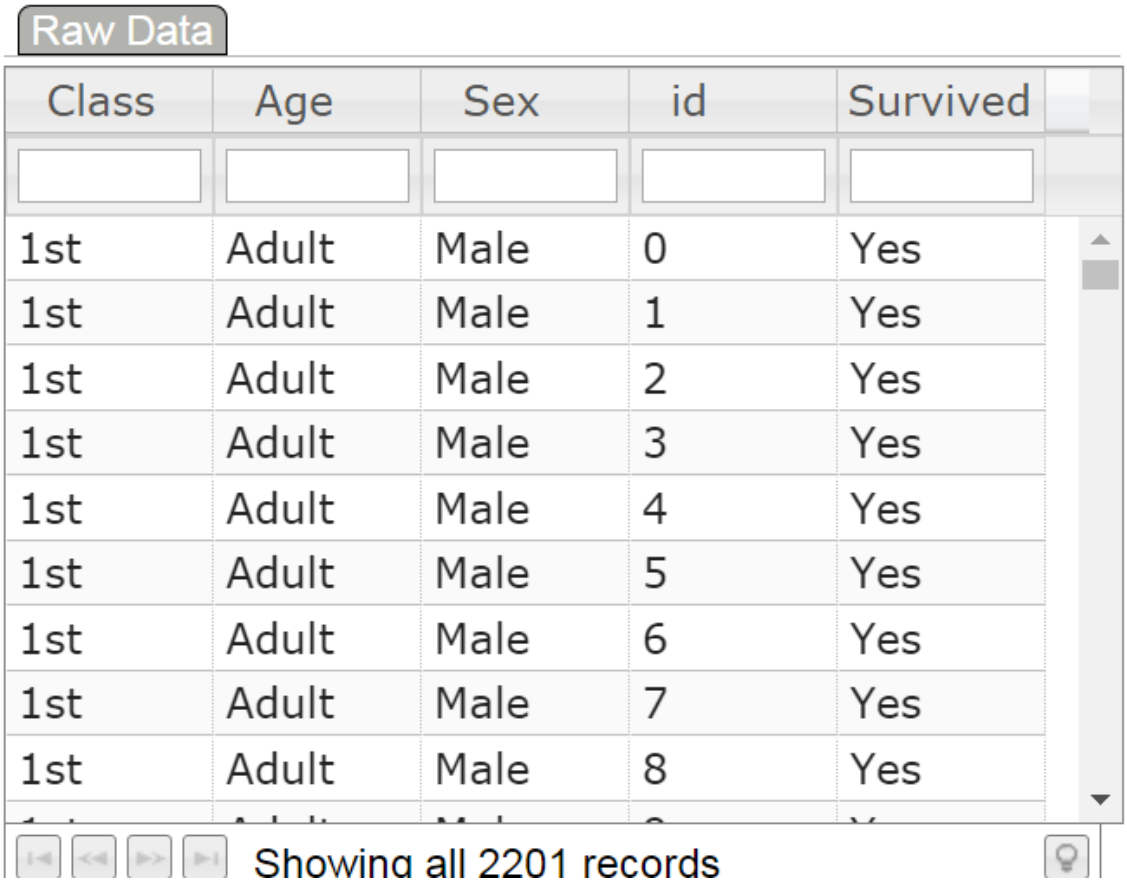

Figure 20: RawData View for The Titanic dataset

the relationship between two categories represented by a ribbon in ParSets. In Figure 21, category *"Child"* in dimension *Age* and category *"No"* in dimension *Survived* are used as filters to search for all children not survived. In this way, users can find out in what circumstances "women and child first" do not apply.

4.12 Other Interactions

Besides sorting, we provide a set of interactions to facilitate interactive visual exploration for variable relationships

Category Reduction To hide the categories that do not exist in any of association rules in the ARTable, users can click a "Show only" button. Figure 19 is before hiding

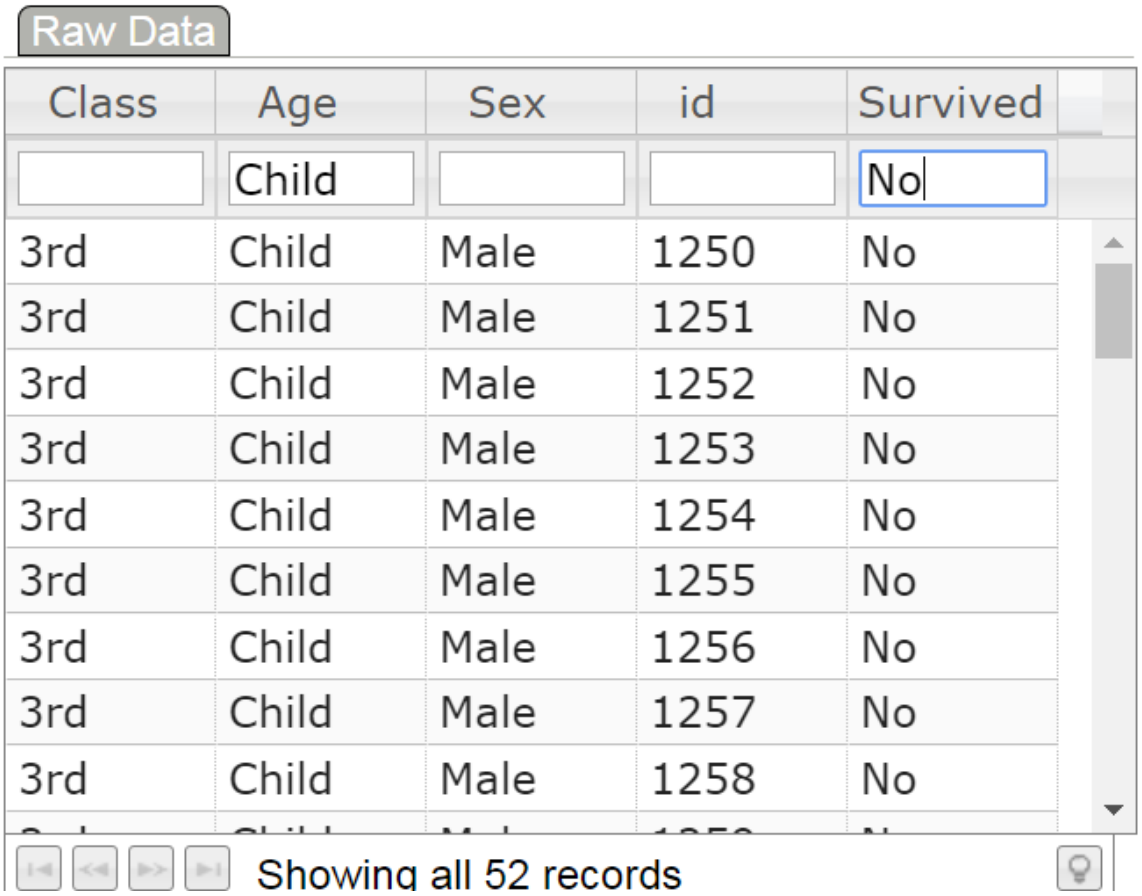

Figure 21: Searching in the Rawdata View. There are 52 child who have not survived.

the categories. Figure 22 shows the display after the categories are hidden. The clutter has been significantly reduced, with the clutter measure changed from 18.89% to 11.49%.

Hovering Over In the ARTable, users can hover the mouse over a row to see the tooltip telling more details about the rule, such as size of rule itemsets, support, confidence, and lift. In ParSets, hovering the mouse over a category will trigger a tooltip showing the absolute numbers and percentages of observations in this category as a fraction of the entire data set. The ribbons connecting this category will be highlighted. When the mouse hovers over a ribbon, the tooltip shows the combination

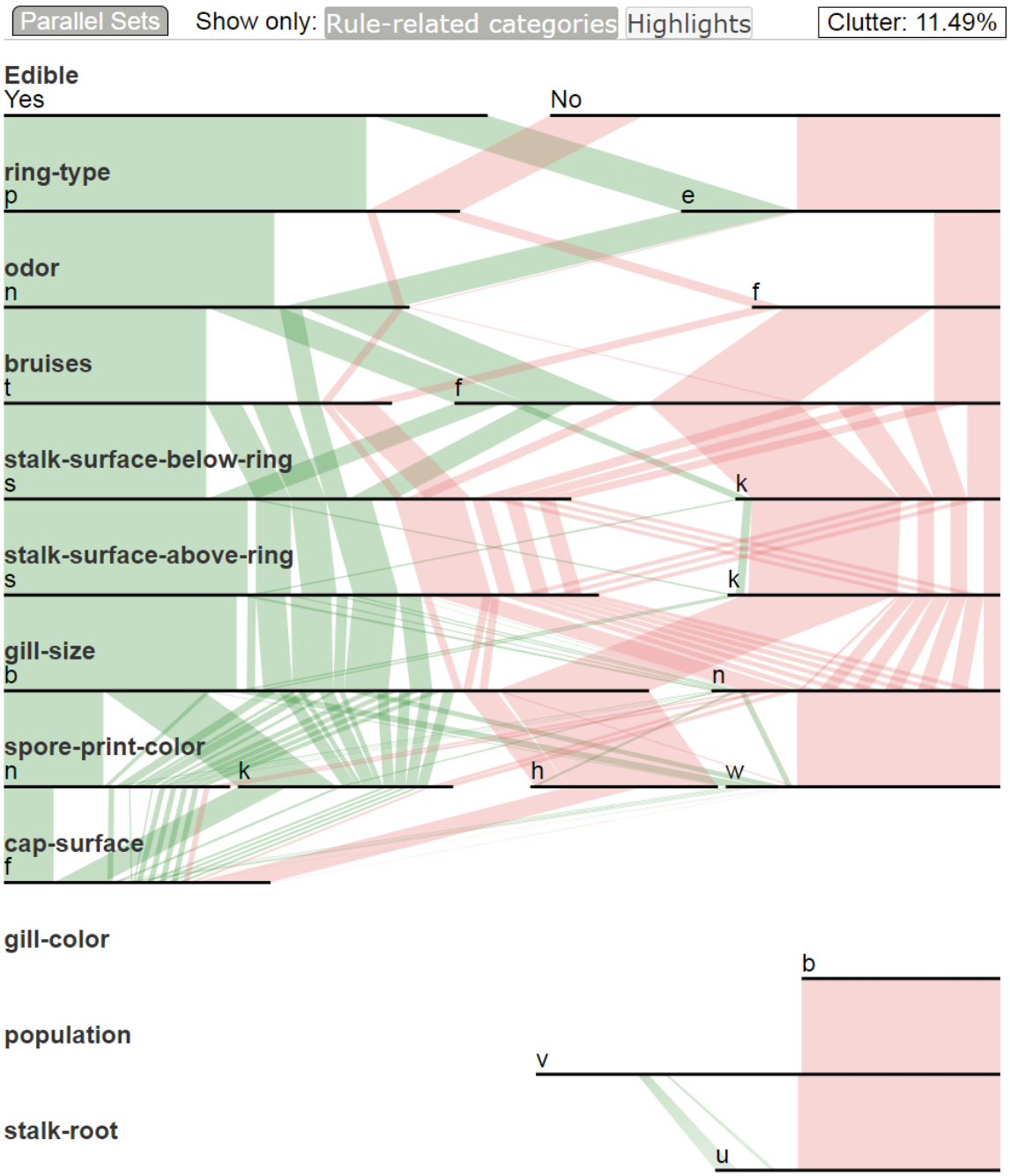

Figure 22: ParSets - Mushroom - *Yes Closeness* - *Confidence* - show only Related Categories. Clutter-reduced ParSets sorting dimensions by *Yes Closeness* and sorting categories by *Confidence* with the only display of rule-related categories. The support range is from 18% to 34%.

of criteria that ribbon represents, as well as absolute and percentages of observations matching that criteria.

Category Selection Clicking a category will make all ribbons connecting to it highlighted, even when the hidden categories are brought back to the view. The rules containing that category will also be highlighted in the ARTable. This interaction enables the following exploration strategy: hiding insignificant categories, clicking a category of interest from the less cluttered view, and then showing hidden categories to examine the selected category within context.

Rule Selection By clicking a rule in the ARTable, we can highlight the set members in ParSets. For example, in Figure 23, the rule containing $\{$*stalk-surface-above-ring* $=$ *s*, *odor* $=$ *n*$\}$ is clicked and is highlighted in yellow border and dark green. Other rules containing the itemsets are also highlighted in dark green.

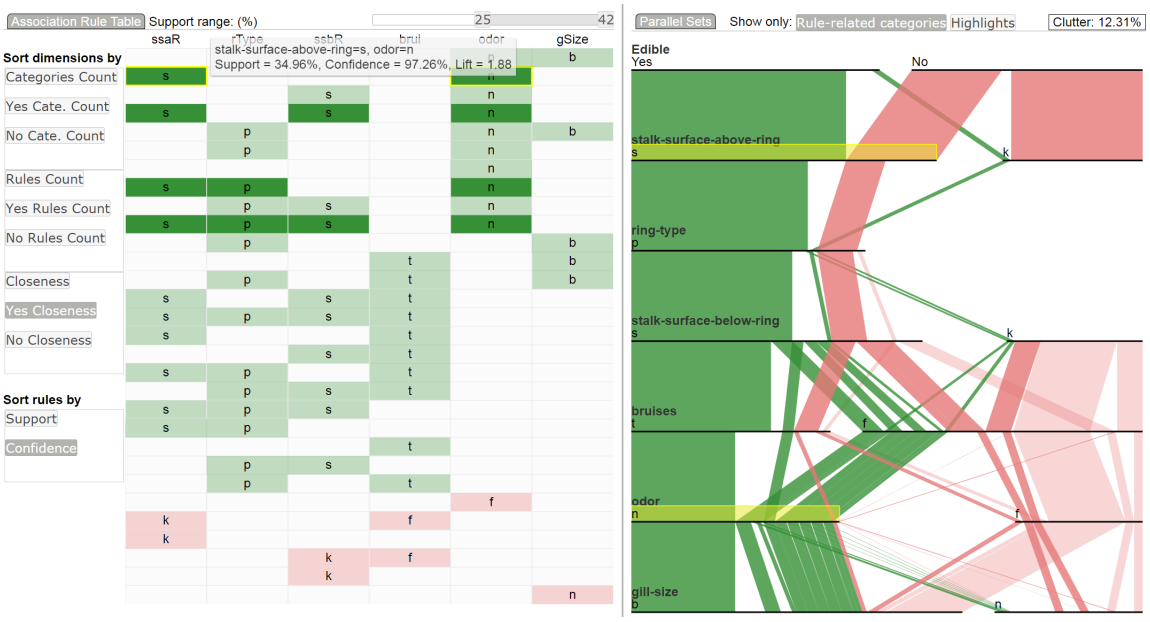

Figure 23: Highlighting on ParSets when clicking on ARtable. Rule containing $\{$*stalk-surface-above-ring* $=$ *s*, *odor* $=$ *n*$\}$ is clicked on the ARTable and is highlighted in yellow border and dark green, the other rules containing the same itemset are also highlighted in dark green. Meanwhile, the corresponding categories are highlighted in the ParSets as well as the ribbons that pass though the dimensions and categories that itemset contains.

Rule Filtering Users can use range filtering on support to change the number of

rules displayed in the ARTable. In particular, they can drag sliders to change the minimal or maximal supports to filter rules. Only rules displayed in the ARTable are used for dimension and category ordering.

Manual Reordering The automated dimension and category ordering approaches are heuristic. We allow users to manually adjust the dimension and category orders for a layout of interest. In particular, we provide drag and drop interactions for ARTable and ParSets. In the ARTable, dimensions can be rearranged by clicking and holding the mouse on the table column and dragging it around. The other categories will move out of its way so that we can place it wherever we want. In ParSets, dimensions and categories can be reordered in a similar way by clicking and dragging a dimension's and category's label.

4.13 Case Study - Finding Characteristics of Edible Mushrooms

The goal of the case study is to find the primary characteristics of edible mushrooms using a clutter-reduced Parallel Sets.

The Mushroom dataset [68] includes descriptions of 23 species of gilled mushrooms which are categorized as either poisonous or edible based on physical attributes. This dataset has multiple categories for some attributes. It contains 8,124 instances and 22 categorical variables. Missing values are replaced with the character *u*.

We use the Apriori algorithm implementation by [11] to generate association rules with the support threshold of 10%, the confidence threshold of 80%, and the lift threshold of 1.2. 117 *Edible* rules and 435 *Poisonous* rules are resulted and displayed in the ARTable.

The 22 dimensions sorted in an alphabetical order are displayed in the ParSets (see Figure **??**). We top the class dimension *Edible* for a better exploration of related characteristics. The clutter is 20.50% with the default support range of 18% to 34%. Now, we pay more attention to high-support rules and change the support range to $\geq$ 25% by dragging the sliders (see Figure 16). Because the outcome of interest is edible mushrooms, we sort dimensions by *Yes* closeness and sort categories by rule confidence to make sure that ribbons representing *edible* characteristics slide to the left of ParSets as much as possible. We receive all reordered dimensions in ParSets that are reflected in ARTable (see Figure 24).

Then we click the "Rule-related categories" to hide categories that are not in the ARTable. The ParSets become more clear (see Figure 25).

By looking at the leftmost categories in Figure 25, there is no single green ribbons across them, which means these categories (characteristics) are not 100% related to edible. Among them, *odor* = *n* has a least wide of red ribbon. We click the category of *n* and all ribbons connecting to it are highlighted recursively (Figure 26).

We wanted to find where the small red ribbon belongs and what other categories are related to it and contribute to being *poisonous* mushrooms. We click the "Rule-related categoris" to go back to the ParSets of the original dataset while keeping the *odor* = *n* highlighted. Figure 27) indicates that the small red ribbon passes through only a few categories in which *spore-print-color* = *r* is the only one occupied red color.

Now, we only click the *Edible Yes*. Figure 28 only shows the highlights of all related categories green. We also click the *Highlights* button and we have Figure 29.

We wanted to know whether the narrow green ribbons have some rules generated.

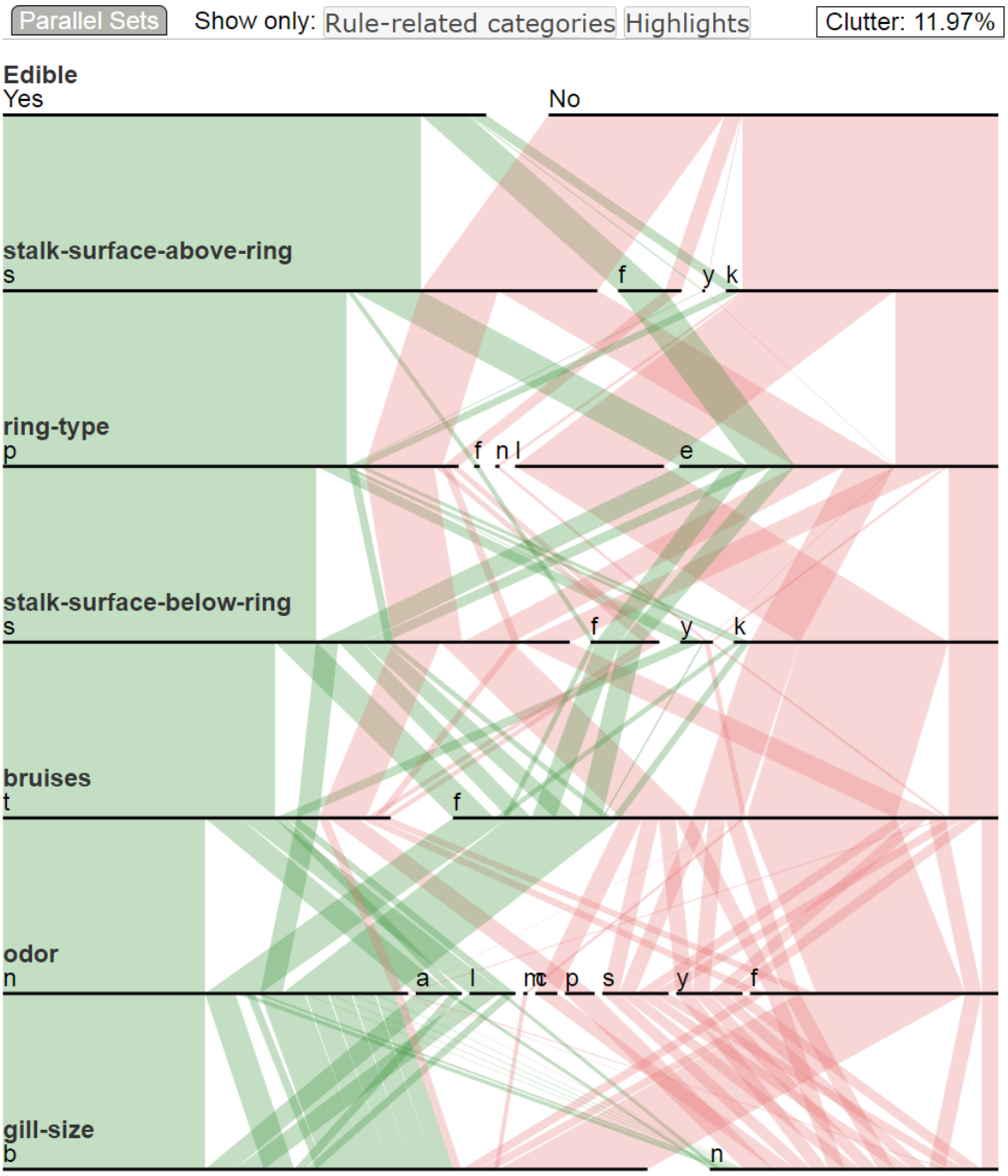

Figure 24: ParSets case study - Mushroom - *Yes Closeness* - *Confidence*. The support range is $\geq 25\%$.

The tooltip shows only 3% observations matched with the ribbon. We click the ribbon and the related rules are highlighted in the ARTable (Figure 30).

We found several articles [25] suggesting that *Odor = n (none)* is a significant

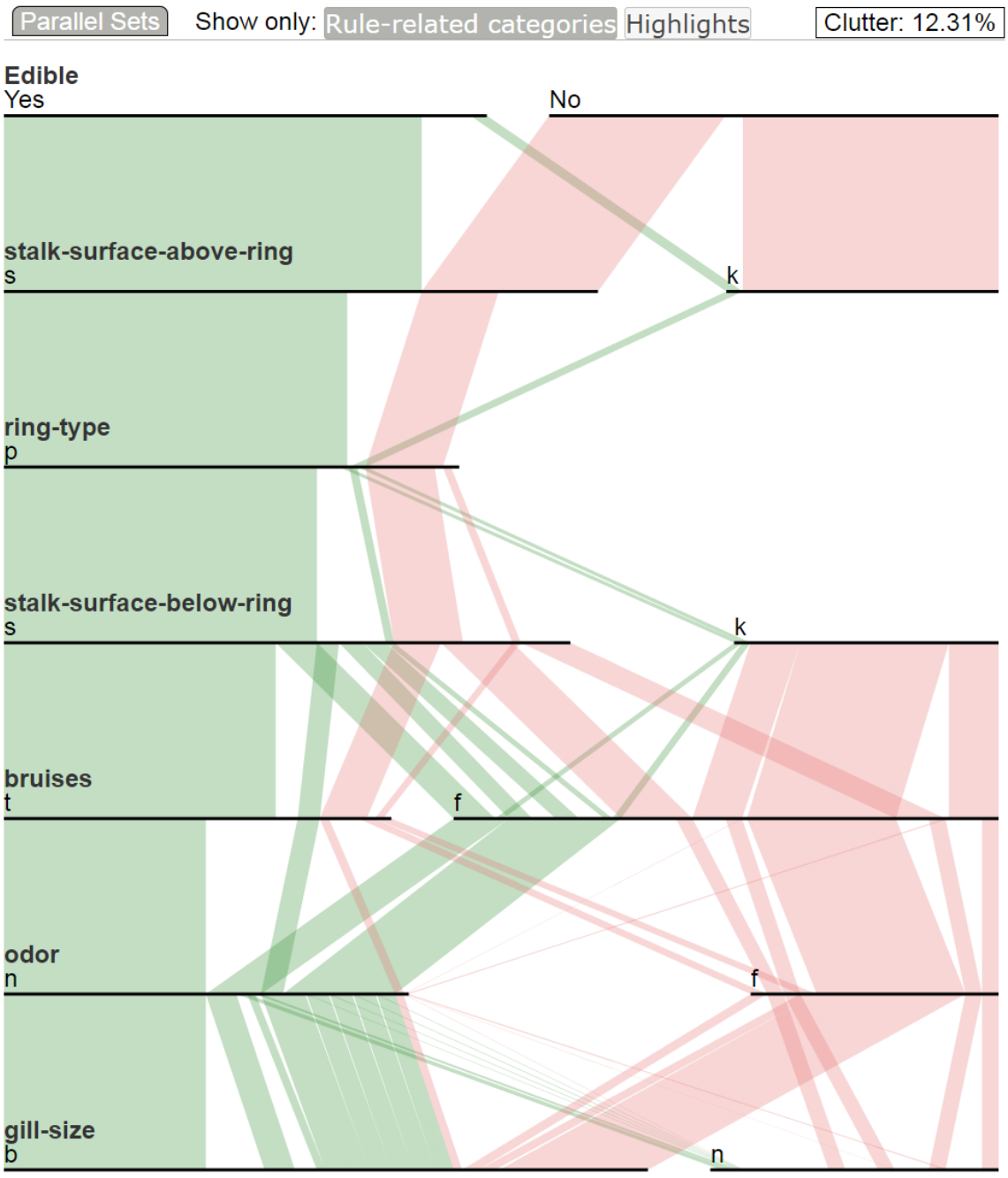

Figure 25: ParSets case study - Mushroom - *Yes Closeness* - *Confidence* - show only Related Categories.

feature for edible mushrooms, which confirmed our finding with the clutter-reduced ParSets.

In this case study, we demonstrated how the clutter-reduced visualization help find

Figure 26: ParSets case study- Mushroom - Category Click. *odor* = *n* has a least wide of red ribbon. Related ribbons are highlighted when the category is clicked.

significant explanatory variables. We also saw clearly the advantages of revealing a small signal in a high noise environment which might not be easily found with cluttered Parallel Sets.

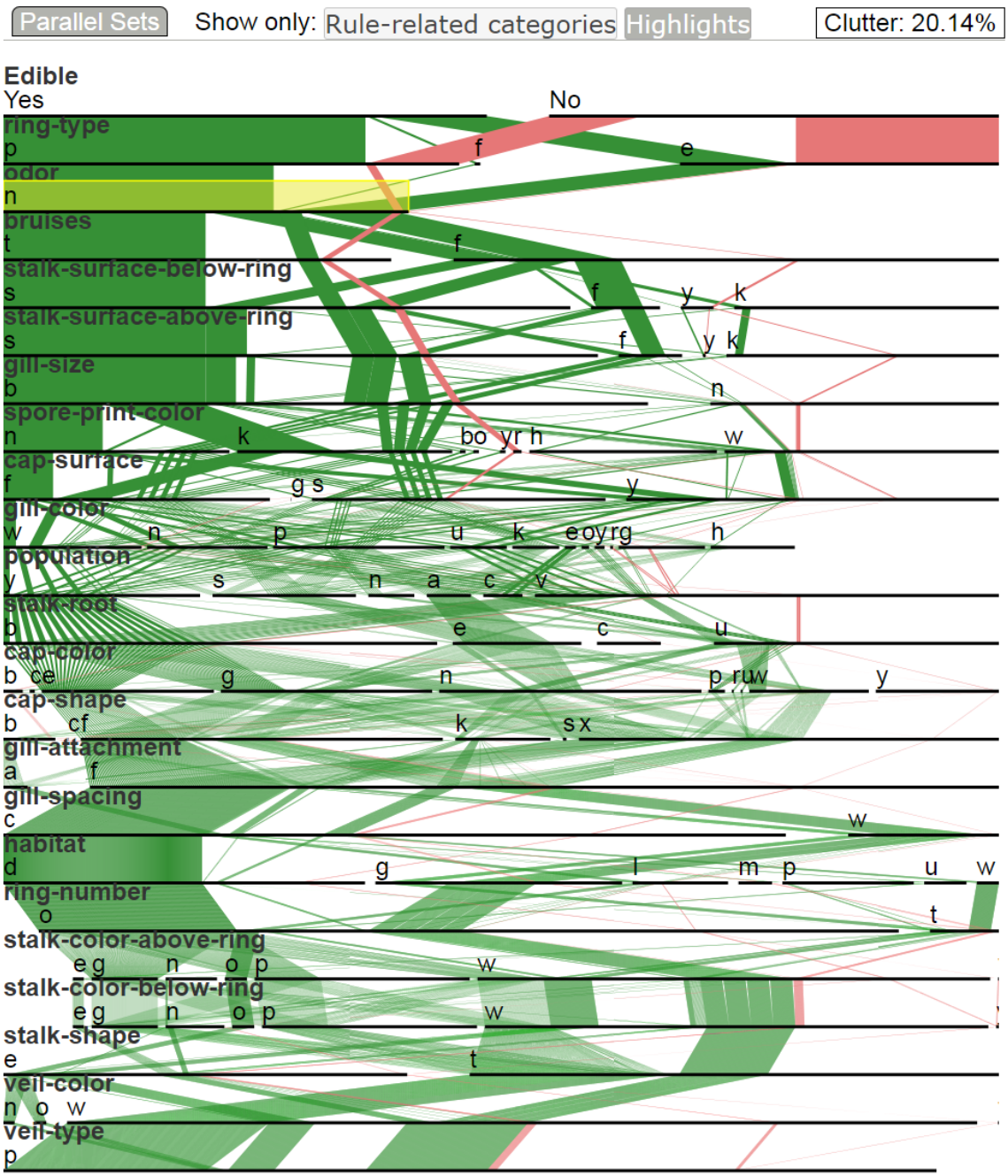

Figure 27: Edible category *odor* = *n* is clicked and highlighted in the ParSets for the original dataset. The small red ribbons through *odor* = *n* mostly pass through *spore-print-color* = *r* which does not have any green ribbons coming in and out.

4.14 Case Study - Rediscovering Characteristics of Caravan Policy Holders

In this case study, we rediscover the characteristics of caravan policy holders using the COIL 2000 datasets, the same one as the use case in Chapter 3.

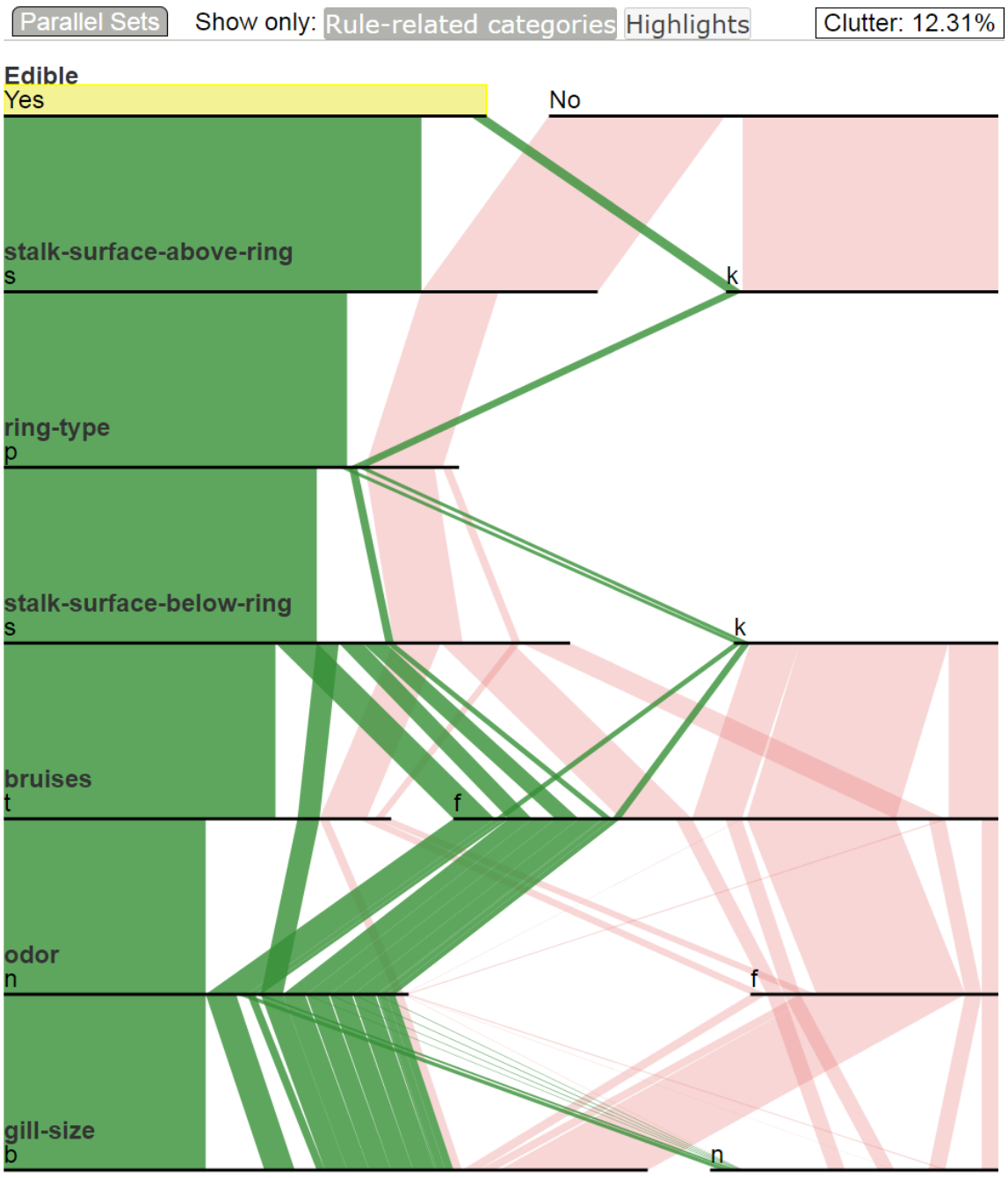

Figure 28: *Edible = Yes* is clicked and related ribbons are highlighted.

The basic idea is that we use the variables selected with the approach proposed in Chapter 3. The subspace dataset is then used as the input for association rule mining. The only difference of variable selection process from chapter 3 at this moment is

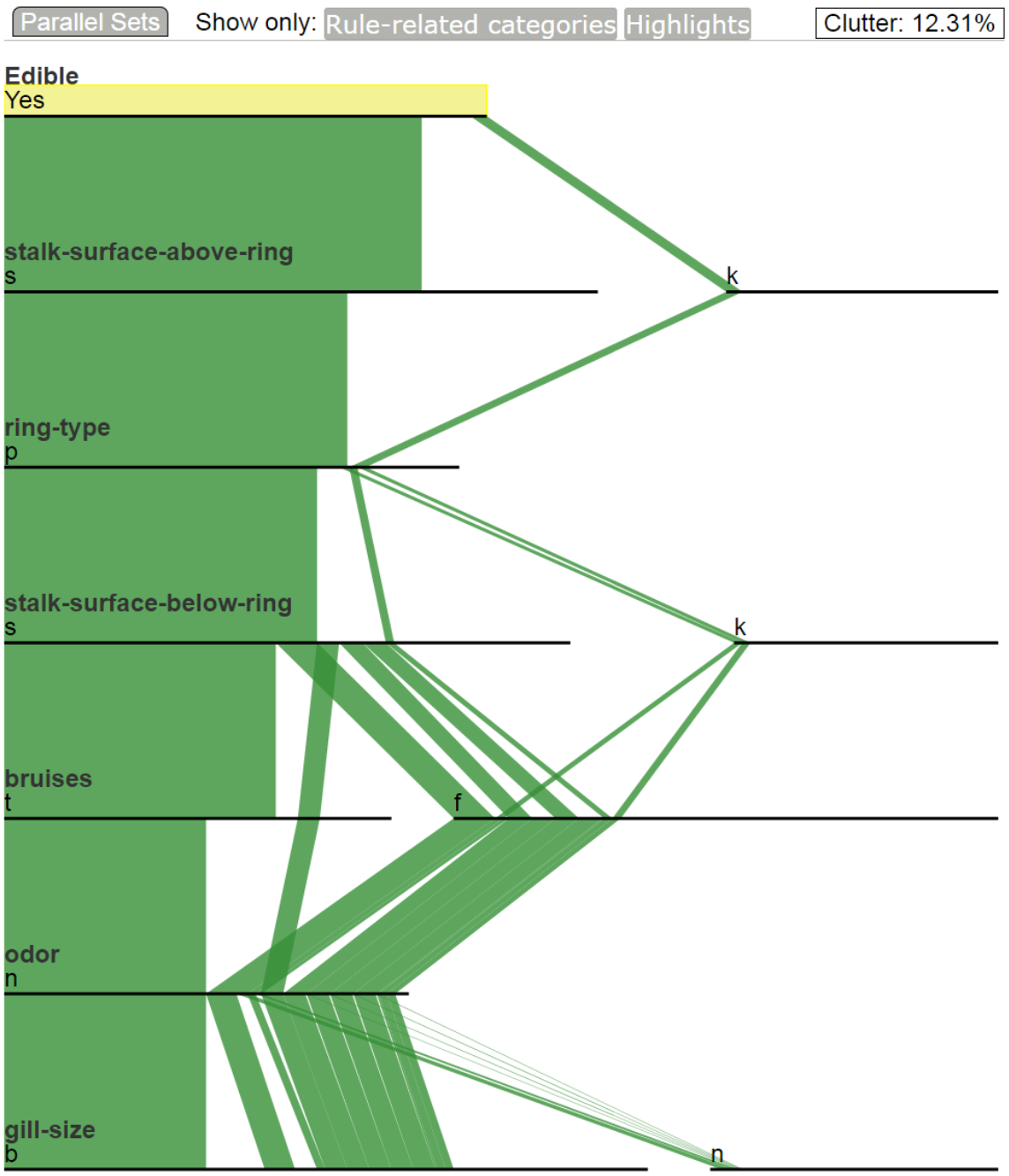

Figure 29: *Edible = Yes* is clicked and related ribbons are highlighted. Only show these highlights.

we consider most variables as continuous. Only *MOSTYPE* and *MOSHOOFD* are considered as categorical because they are nominal.

We tolerate confounding at this time, stop at the first iteration of variable selec-

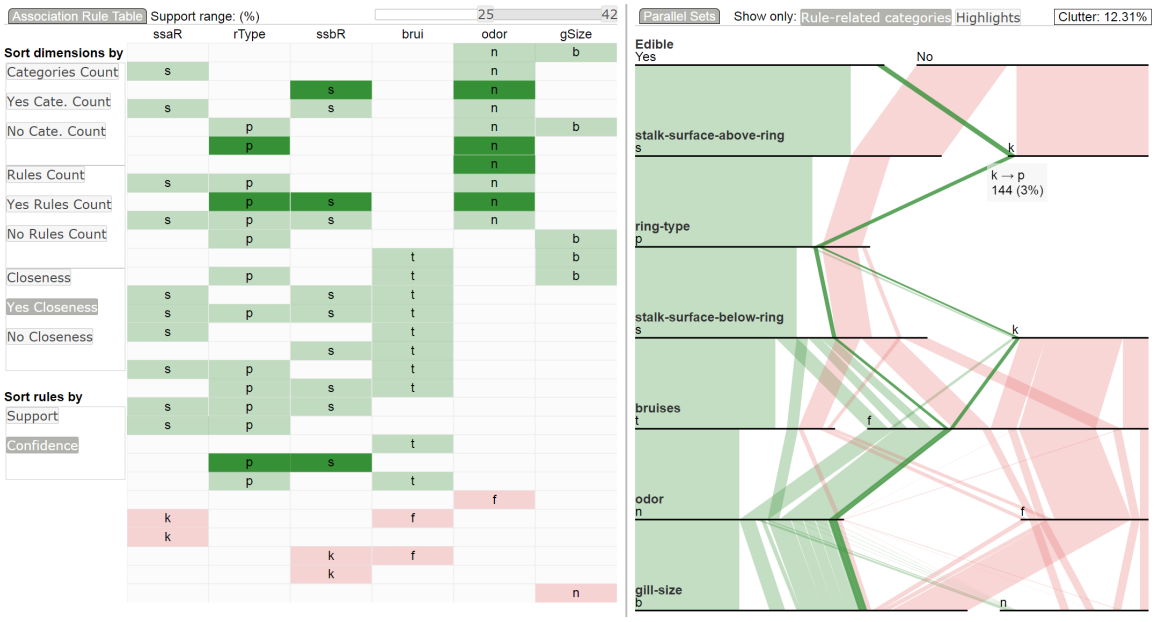

Figure 30: ParSets case study- outlier ribbon inspection

tion, and select all variables with yellow tag in the Variable Group View (Figure 31). This is equivalent to saying that we only focus on the significant variables presented in the *Univaraite* column in the Model Evaluation View. *MOSTYPE, MOSHOOFD, MRELGE, MGODPR, PPERSAUT, MBERMIDD, MKOOPKLA, MOPLHOOG, AFIETS, MFWEKIND, MBERHOOG, MINKGEM, and PBYSTAND* are used as input variables for association rule mining.

The caravan policy holders are rare in this dataset. A very small percentage of people have the policy (586 out of 9,822). We set the *Yes* class to have the support threshold of 1% and *No* class 10%. There are 109 *No* rules and 25 *Yes* rules generated. The maximum support of *Yes* rule is less than 5%. Since we are only interested in the *Yes* rule, we set the support range to 0-5 (%) and use *Yes Closeness* and *Confidence* to order dimensions and categories (see Figure 32).

We click the *CARAVAN Yes* label to highlight all the related categories. We can

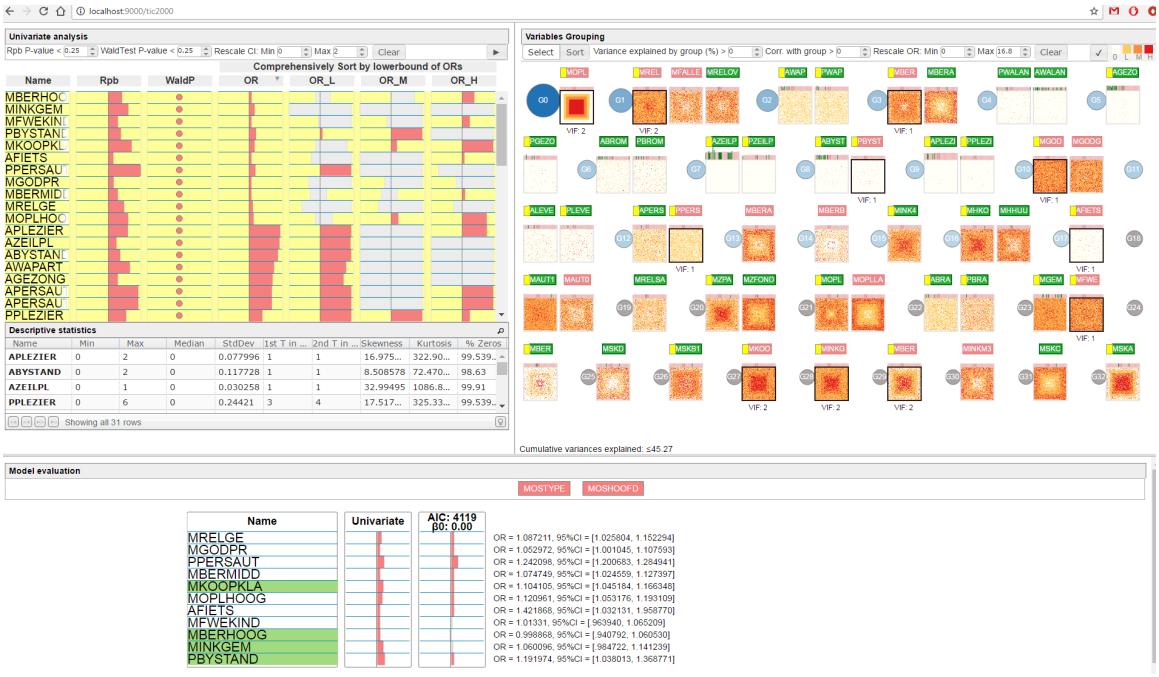

Figure 31: Variable selection process for the COIL 2000 dataset using the proposed approach in Chapter 3.

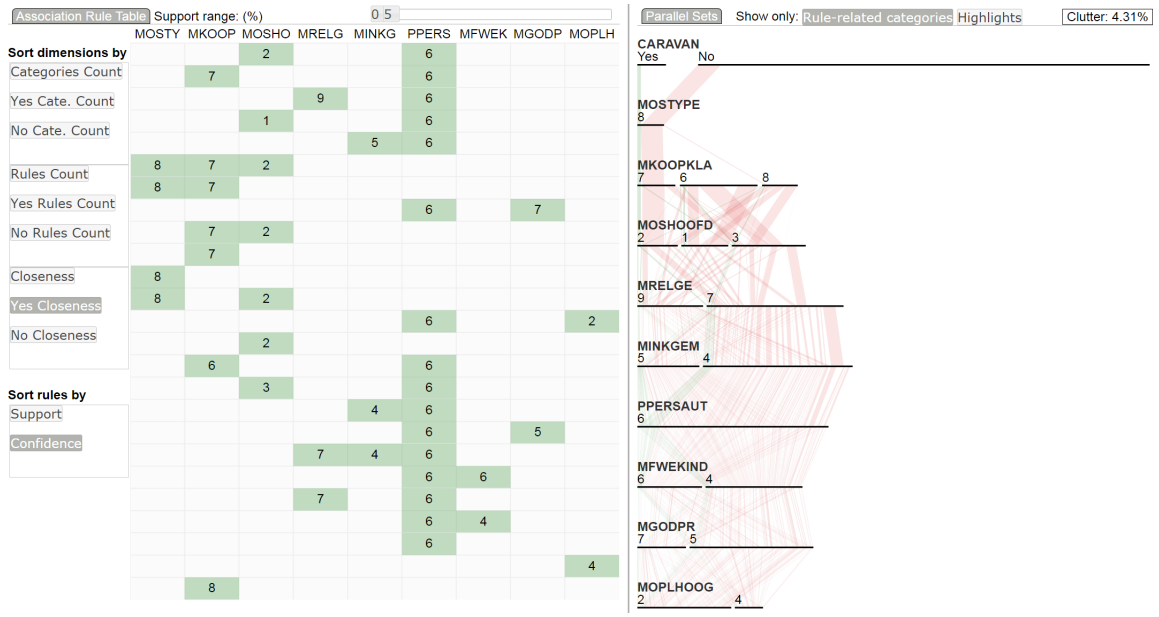

Figure 32: Rule selection for COIL 2000 dataset in ARTable. The support range is 0 - 5(%). *Yes Closeness* and *Confidence* are used to order dimensions and categories. Most green ribbons are positioned in the left of ParSets.

confirm the small percentage of caravan policy holder by looking at the small green ribbon as shown in Figure 33.

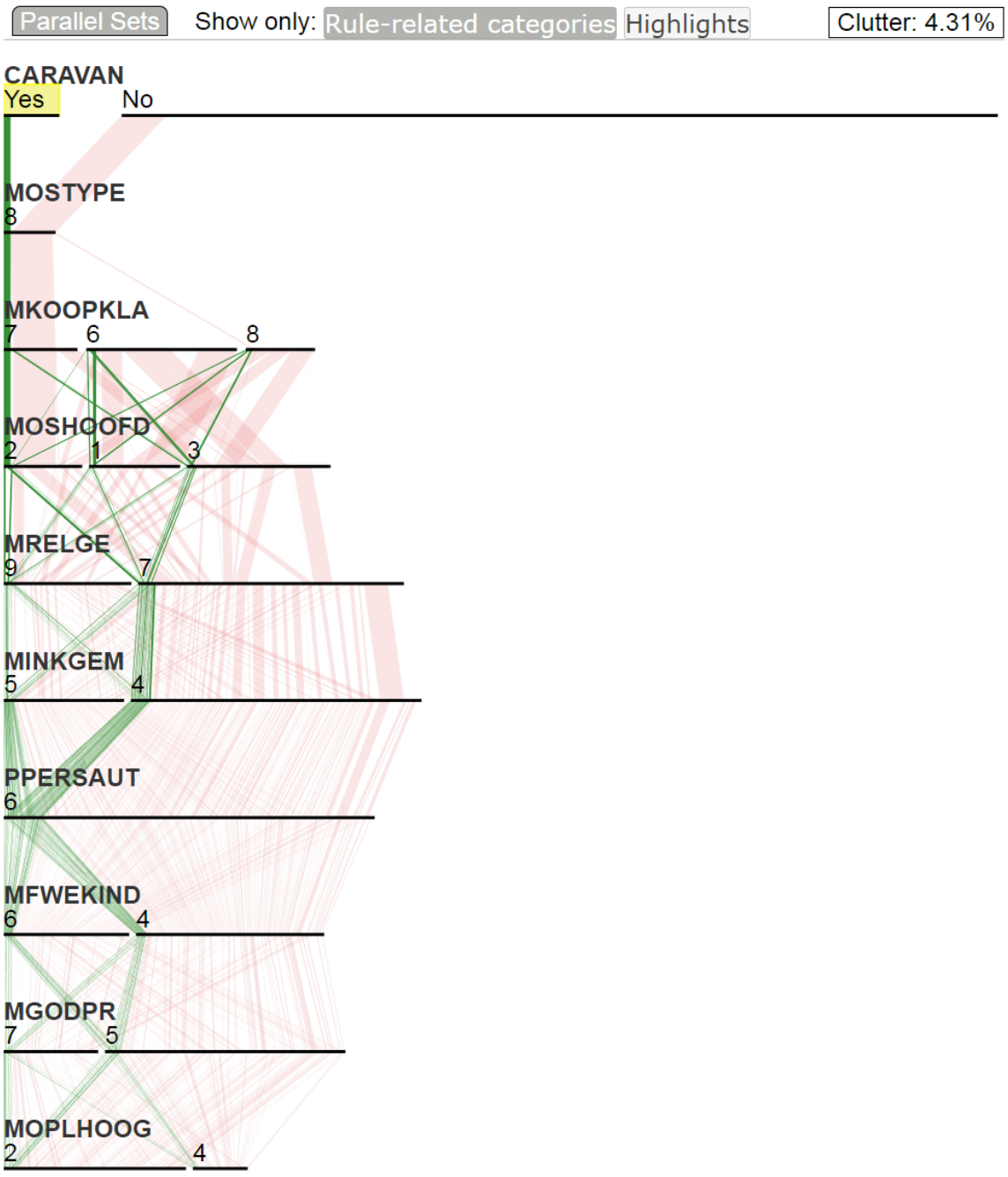

Figure 33: *Caravan = Yes* is clicked and related categories are highlighted. *Yes Closeness* and *Confidence* are used for ordering.

We can see from Figure 33 that the leftmost ribbon is more likely to describe the main characteristics. But after the ribbon passes through $MKOOPKLA = 7$, it splits into smaller ribbons and then merges to $PPERSAUT = 6$. The dimension *PPER-*

SAUT has only one category of *6* displayed, so we change the dimension ordering to by *Yes Cate. Count*. We receive Figure 34.

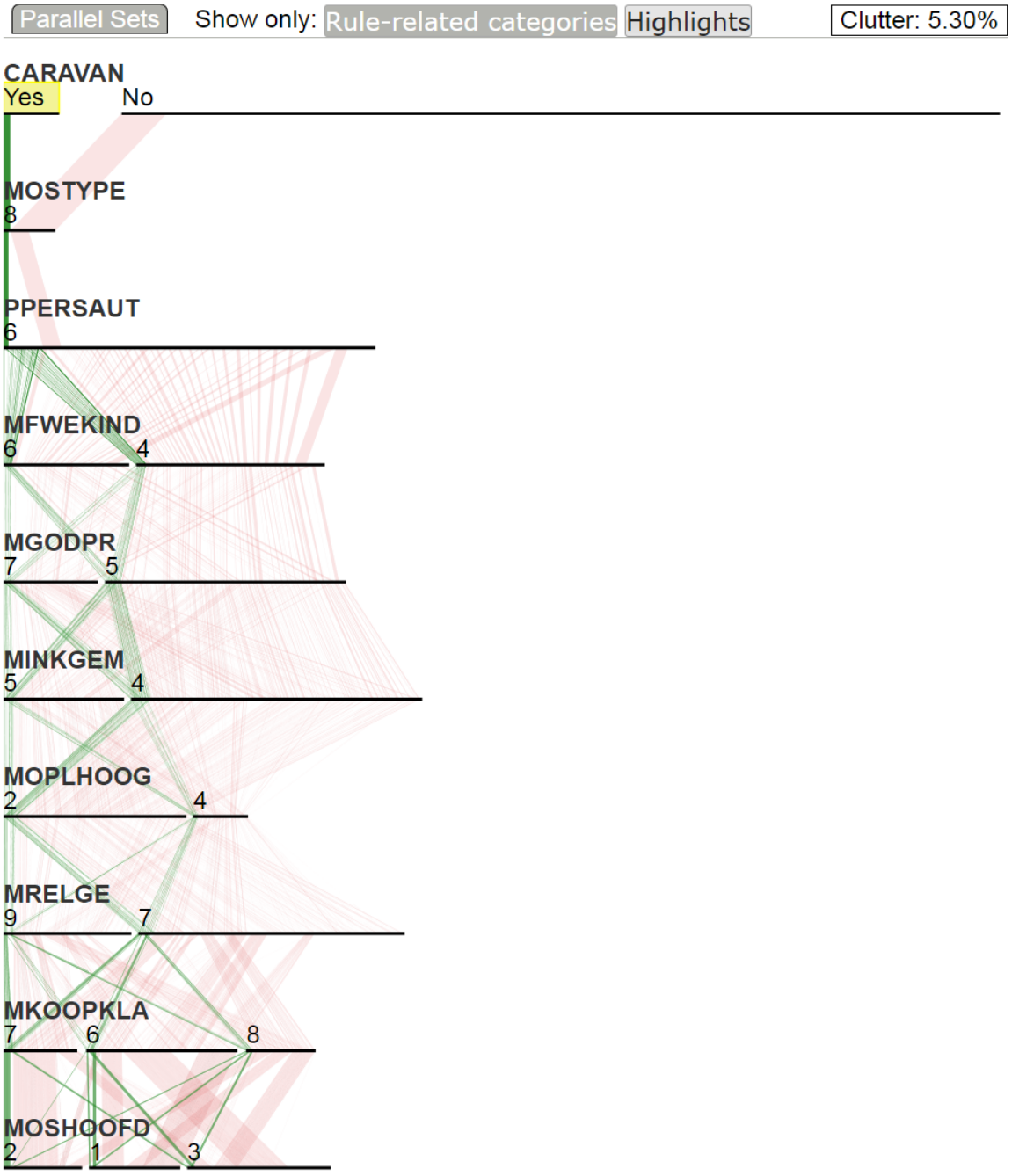

Figure 34: *Caravan = Yes* is clicked and related categories are highlighted. *Yes Cate. Count* and *Confidence* are used for ordering.

There is still a wider ribbon in the bottom dimension *MOSHOOFD* in Figure 34

so we consider ordering these dimension using *Yes Rule Count* (see Figure 35).

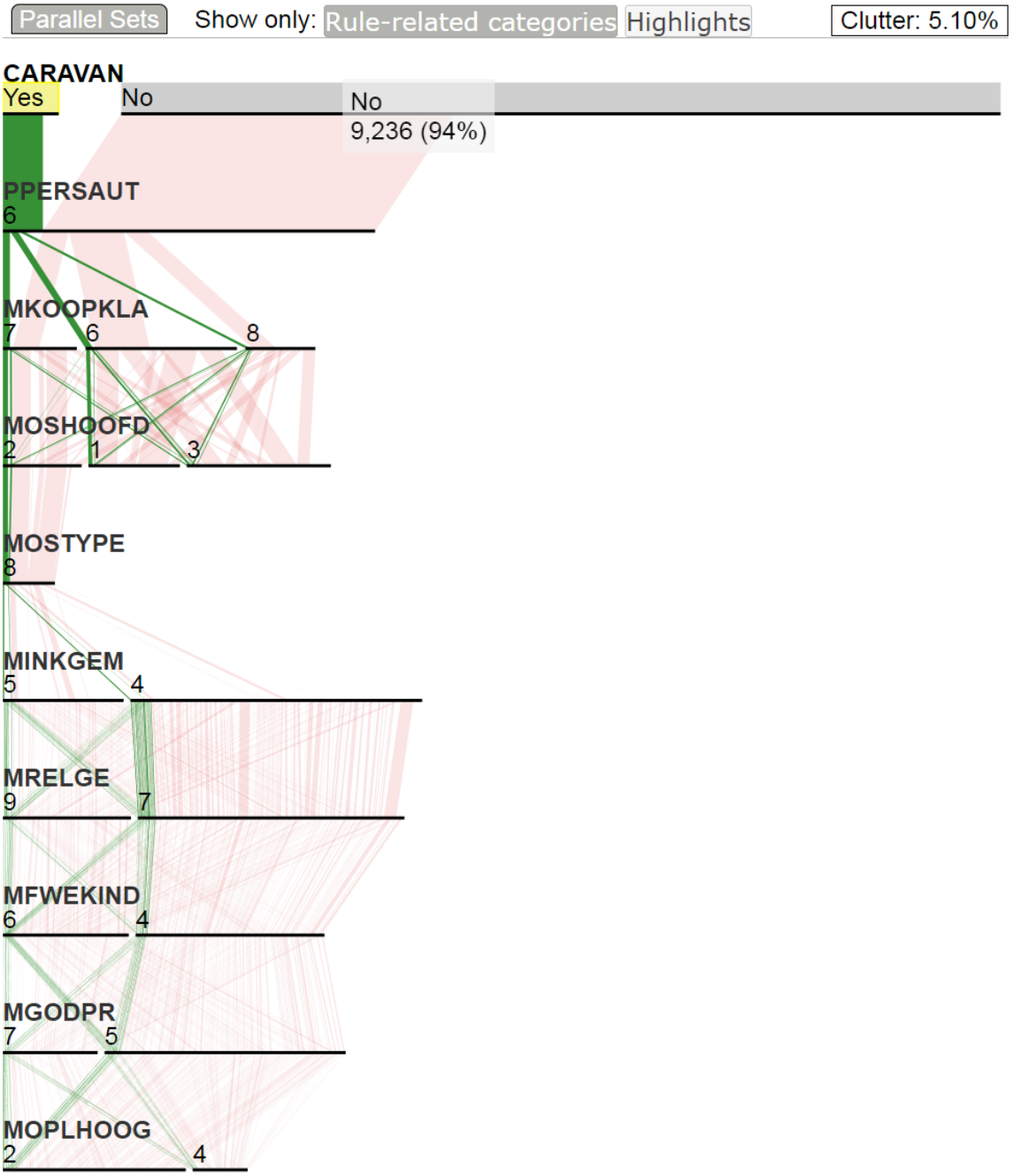

Figure 35: *Caravan = Yes* is clicked and related categories are highlighted. *Yes Rule Count* and *Confidence* are used for ordering.

Now it seems like all wider ribbons are closed to each other. So we can summarize the main characteristics:

- PPERSAUT (Contribution car policies) = 6, range is [0, 9]
- MKOOPKLA (Purchasing power class) = 7, range is [1, 8]
- MOSHOOFD (Customer main type) = 2 (Driven Growers)
- MOSTYPE (Customer Subtype) = 8 (Middle class families)

Although we used the same dataset and had a same purpose in the two case studies discussed in this chapter and Chapter 3, they have different theories no matter what findings could be similar. On the basis of regression modeling, the approach in Chapter 3 considers variable as a whole and provide an understanding of how much the outcome variable changes for one-unit change in explanatory variable. This chapter is built on the co-occurrence association and does not consider the change trend.

4.15 Experiment with Benchmark Datasets

The experiment study compares the effectiveness of clutter reduction with dimension ordering in Parallel Sets for several benchmark datasets in the UCI Machine Learning Repository [68]. We test the *Mushroom* dataset and the *Congressional Voting Records* dataset which are the same as the one used in the article [4]. We would like to find how the dimension ordering methods we proposed work on these different datasets for clutter reduction, how the association rule mining helps create the ordering, and whether the outcome of interest affects the result.

We elaborate the experiment with the Mushroom dataset [68] as an example. For other datasets, we use the same process. The experiment results are summarized in Table 2.

We use different dimension and category sorting algorithms to examine which results in a better clutter reduction for finding the related characteristics of edible and poisonous mushrooms. For the dimension closeness sorting, we use Average Linkage (see chapter 4.9.1). We consider the recent entropy-based dimension ordering approach [4] as one of the benchmarks. Figure 37 shows the ParSets using optimized mutual information ordering for dimension and optimized joint entropy ordering for categories.

As shown in Figure 37, the clutter is 29.72%. However, the alphabetical ordering for dimensions and categories results in a clutter reduction of 20.50% shown in Figure **??**. We uses the same dataset (the original data without removing instances having a missing value), the same browser, the same screen resolution.

Now, we explore the benefit of association rule in terms of the clutter reduction. We first filter these rules using a support range from 25% to 40%. Figure 40, 44, 44 present the comparison of the clutter reduction with sorting dimension by

- Dc: category count,
- Dr: rule count, and
- Ds: closeness

The categories are in alphabetical order for each dimension in Dc, Dr, and Ds. The clutter metrics are 17.57%, 20.29%, and 23.39%, respectively.

We add the rule confidence for category ordering to see if it can improve the above Dc, Dr, and Ds. The results are respective Dc-cc, Dr-cc, and Ds-cc as shown in Figures 41, 43, 45. The measured clutters are 9.63%, 10.82%, and 12.31%.

Next, we set the focus is on the edible mushroom and find the related characteristics. We sort dimensions by

- YDc: Cate Yes Count,
- YDr: Rule Yes Count, and
- YDs: Yes Closeness.

Since we want to find significant edible categories that show on the left side as many as possible, we use rule confidence to sort the categories for all dimension orders above. As shown in Figures 46, 47, 48, the clutters become 9.62%, 10.82%, and 10.32%.

If the focus is on the characteristics of poisonous mushrooms, the three sorting algorithms have a different strength in terms of the clutter reduction (see Figures 49, 50, and 51). Figure 51 shows the clutter is 8.49% and describes the ordering of No closeness for dimension and rule confidence for category with the focus of poisonous characteristics, which has reduced the clutter most.

The experiment result was summarized into Table 2. The first four ordering are the methods proposed in [4]. Among of them, the ones ending in (R) means that the dimensions have been reduced using association rule mining which has exactly the same dimensions as our approaches. Bold numbers indicate the minimum of clutter metrics. We conclude that the effectiveness of clutter reduction is a subjective to the ordering algorithms. A good use of sorting algorithms depends heavily on the dataset to be analyzed and visualized, the significant association rules extracted, as

well as the outcome we focus on. There is no a particular ordering algorithm that works best for all categorical datasets and the resulting association rules. Overall, the association rule tends to have benefits to dimension and category ordering and clutter reduction for a better pattern discovering.

Table 2: Comparison of clutter reduction of different ordering for benchmark datasets. Bold numbers indicate the minimum of clutter metrics.

Dimension sorting	**Category sorting**	**Clutter measure of Mushroom supp.(%) ≥25**	**Clutter measure of Voting supp.(%) ≥36**
Optimal Mutual Information	Alphabet	25.62%	24.39%
Optimal Mutual Information	Joint Entropy	29.72%	15.38%
Optimal Mutual Information(R)	Alphabet	21.30%	25.86%
Optimal Mutual Information(R)	Joint Entropy	11.48%	8.43%
Category count (both classes)	Alphabet	21.32%	23.47%
Category count (both classes)	Confidence	10.10%	8.28%
Rule count (both classes)	Alphabet	23.78%	31.06%
Rule count (both classes)	Confidence	10.96%	8.48%
Association-based (both classes)	Alphabet	26.55%	28.40%
Association-based (both classes)	Confidence	11.97%	8.27%
Category count (class 1 only)	Confidence	10.10%	8.28%
Rule count (class 1 only)	Confidence	10.96%	8.48%
Association-based (class 1 only)	Confidence	11.97%	**8.06%**
Category count (class 2 only)	Confidence	10.10%	8.48%
Rule count (class 2 only)	Confidence	10.96%	8.48%
Association-based (class 2 only)	Confidence	**9.53%**	8.27%

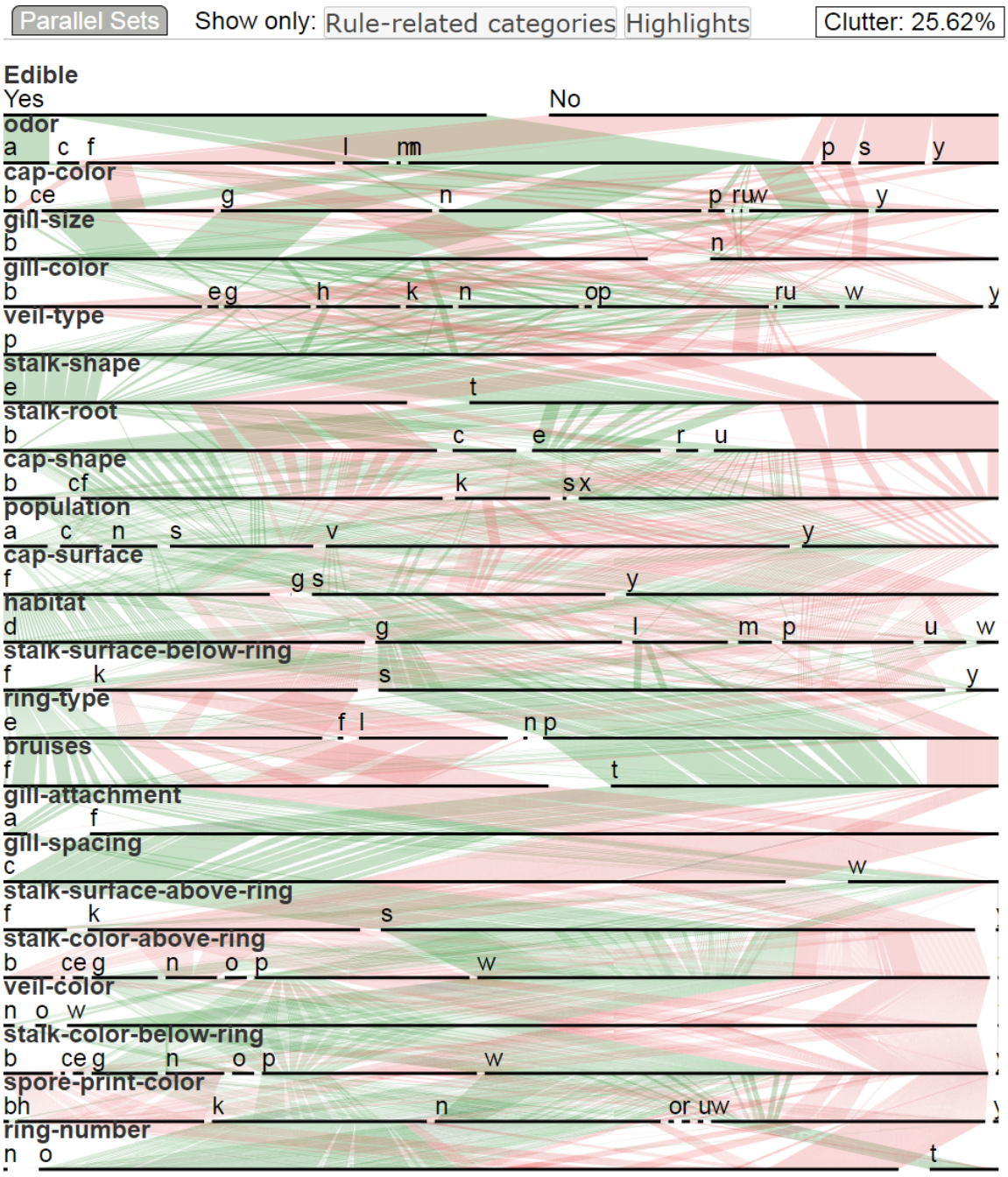

Figure 36: Clutter-reduced ParSets optimized by mutual information of dimensions and alphabetical order of categories, adapted from Alsakran et al. [4]

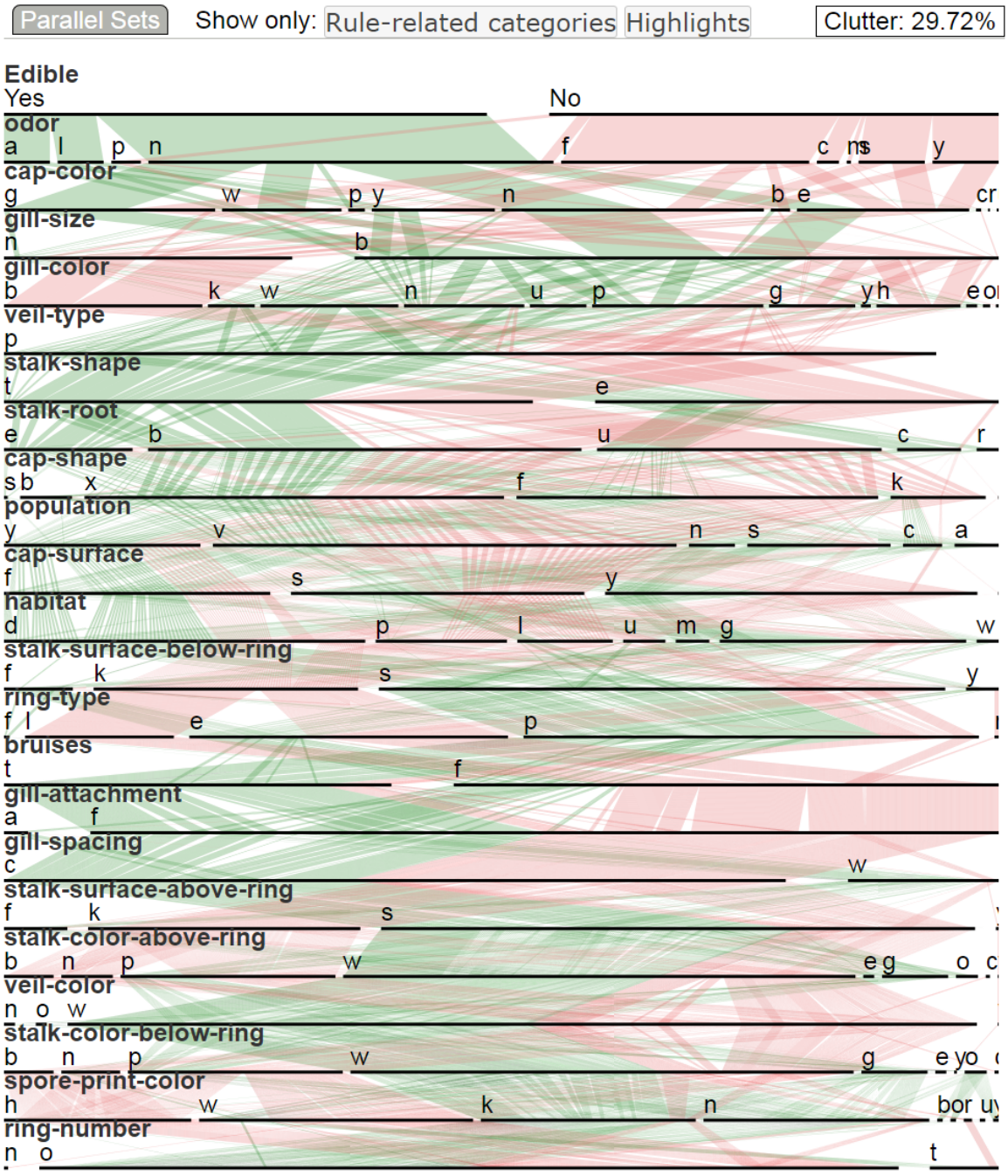

Figure 37: Clutter-reduced ParSets optimized by mutual information and joint entropy.

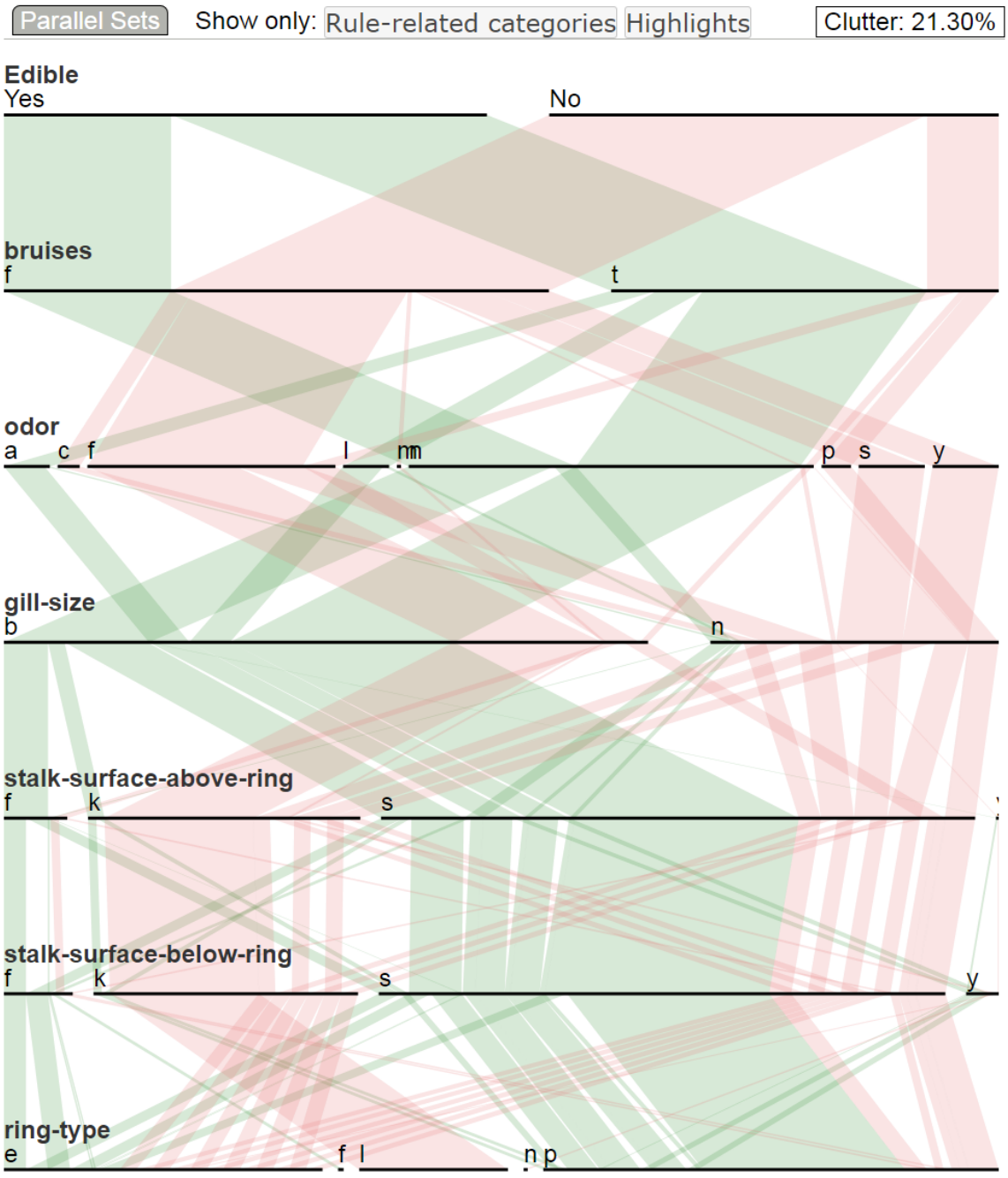

Figure 38: Clutter-reduced ParSets optimized by mutual information of dimensions and alphabetical order of categories. Dimensions are reduced using association rules.

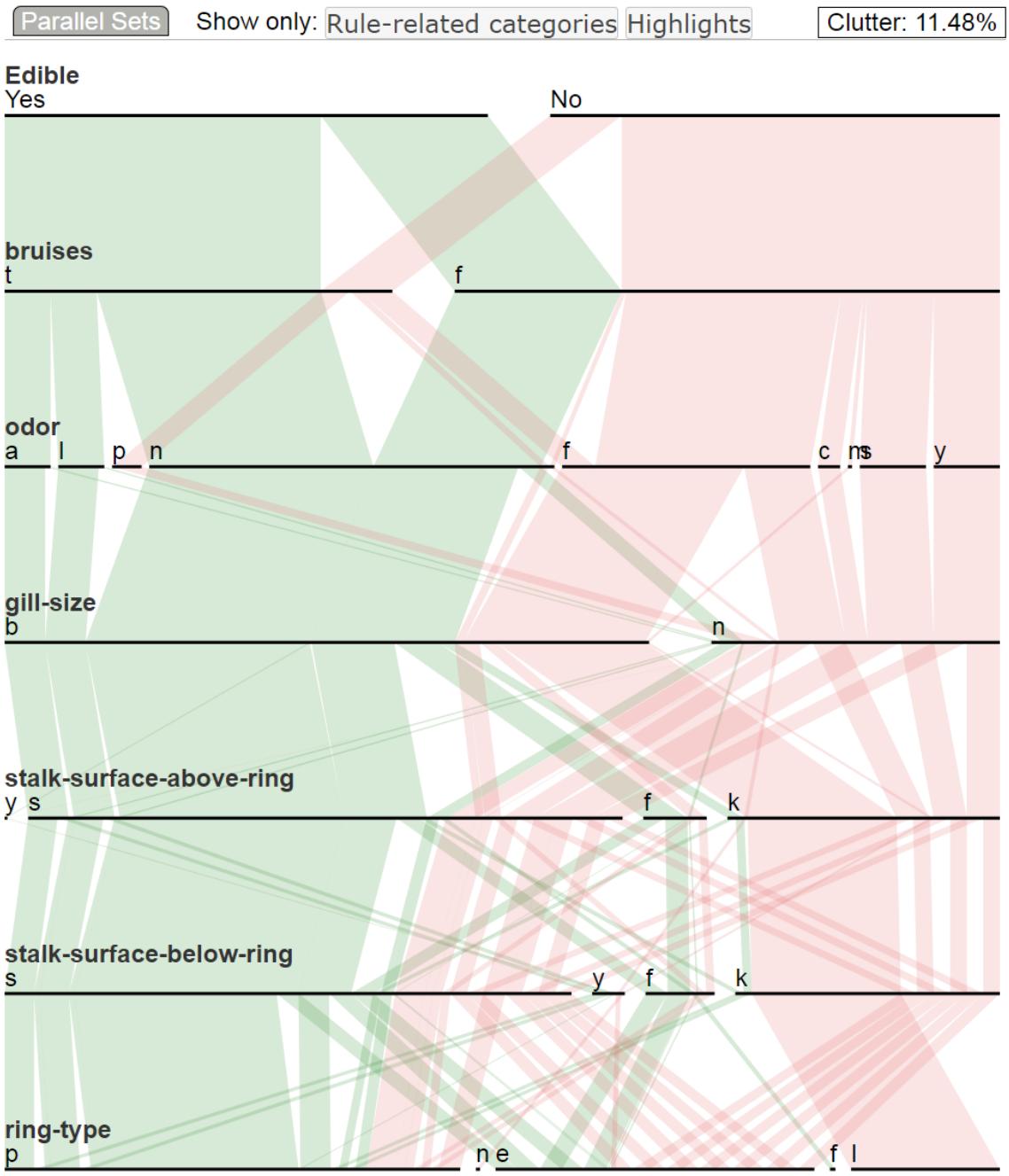

Figure 39: Clutter-reduced ParSets optimized by mutual information and joint entropy. Dimensions are reduced using association rules.

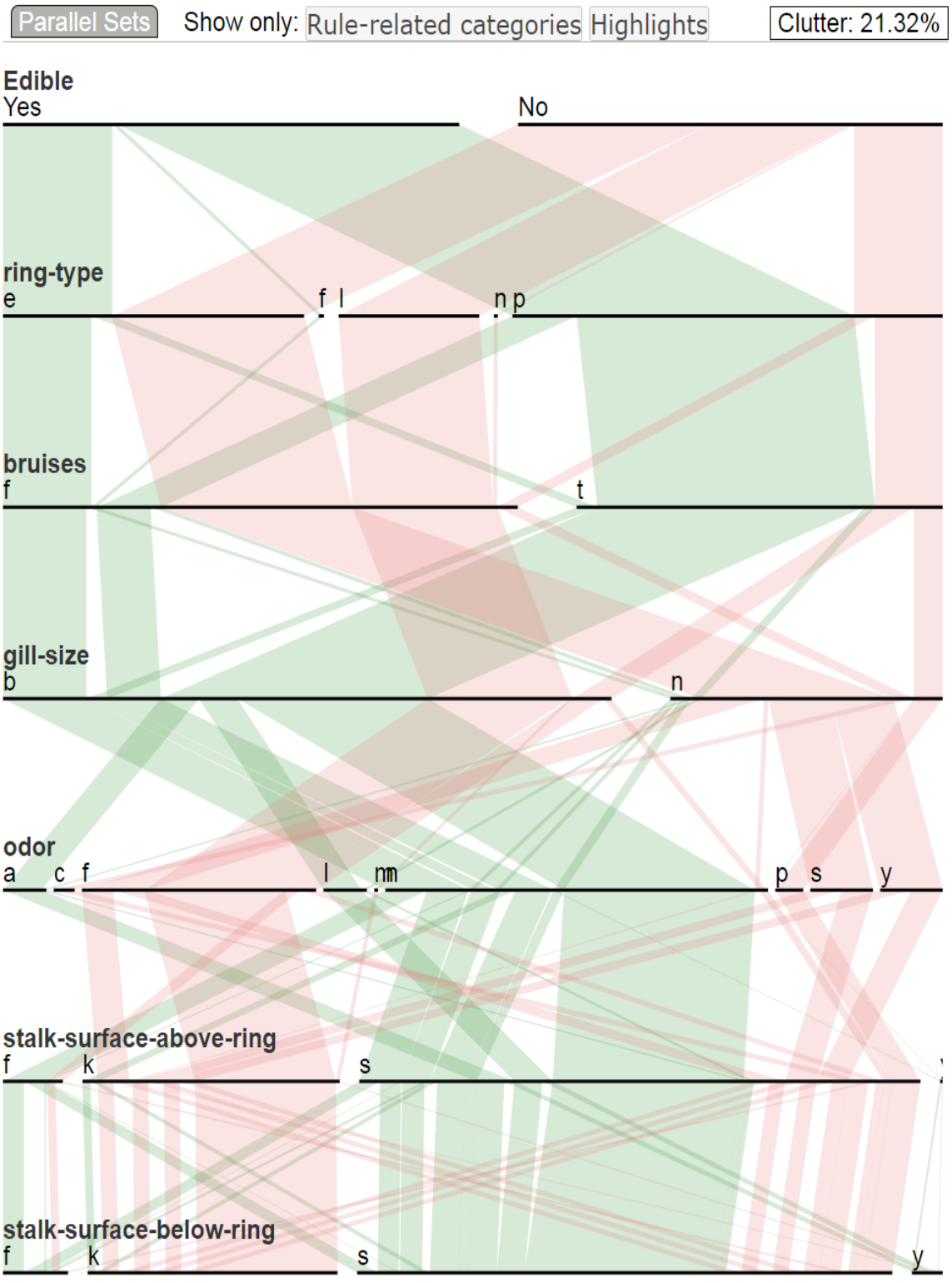

Figure 40: Experiment of Mushroom dimension ordering using Dc (Category Count) for dimensions and alphabetical order for categories.

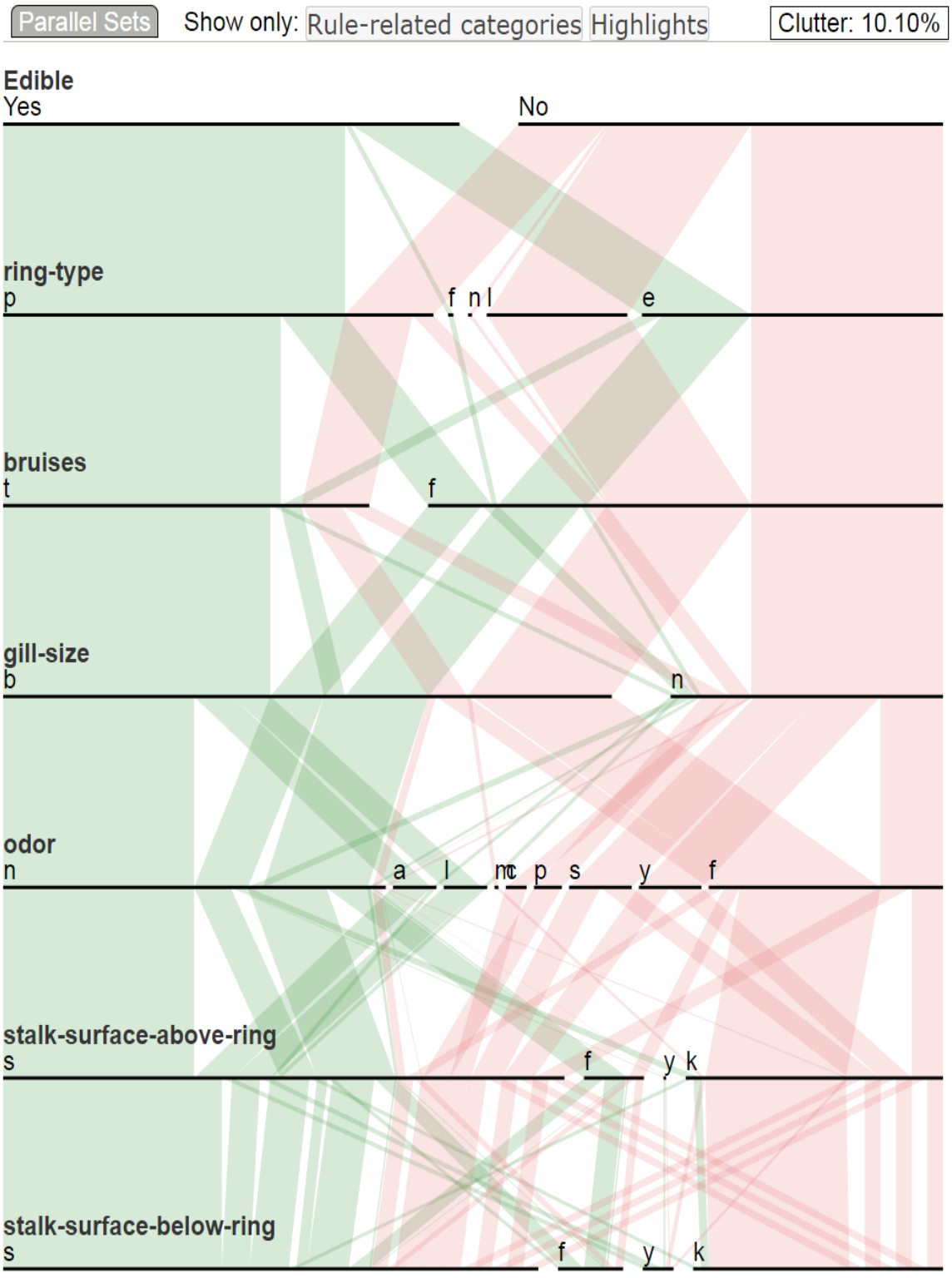

Figure 41: Experiment of Mushroom dimension ordering using Dc (Category Count) for dimensions and Confidence for categories.

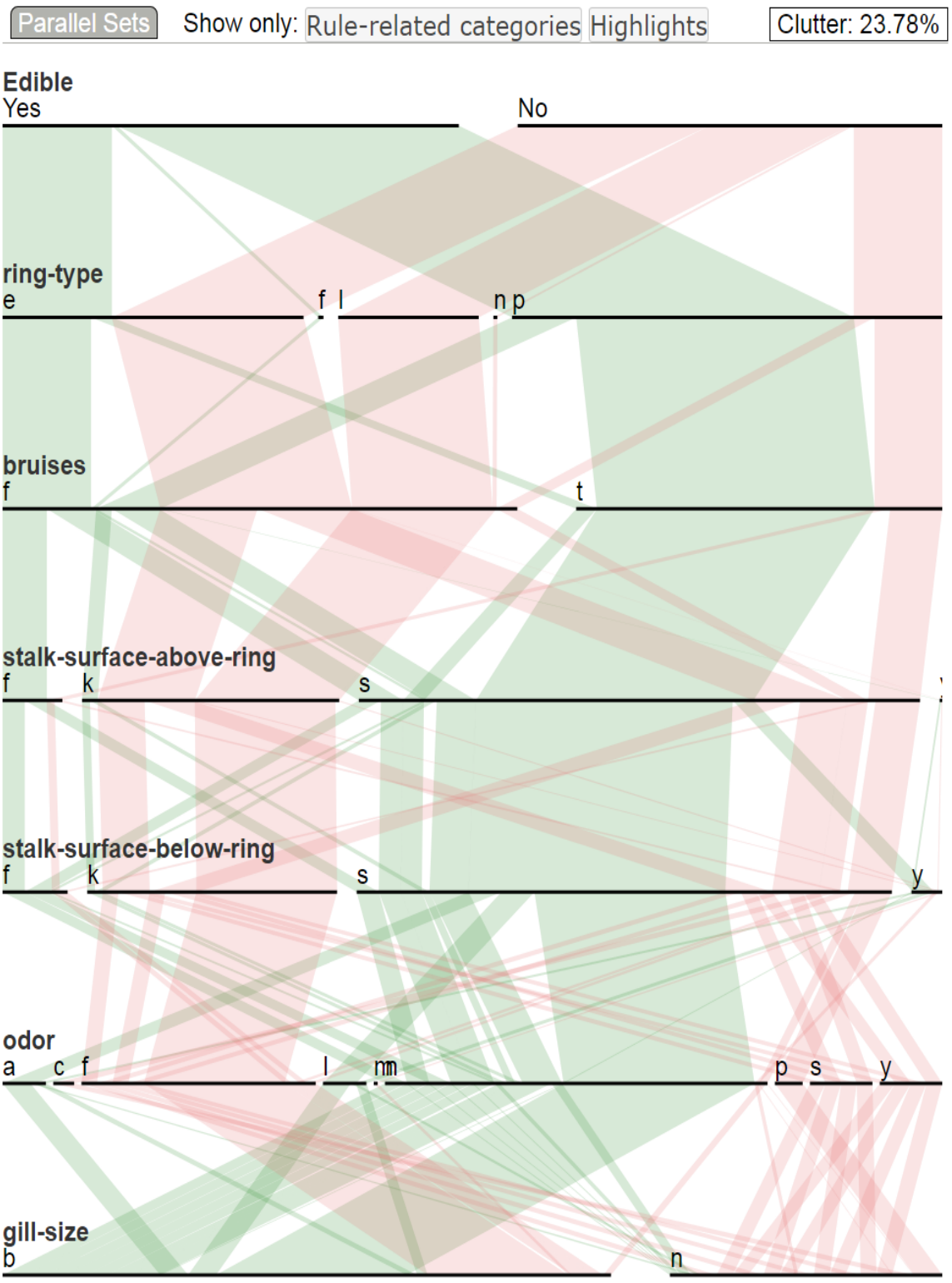

Figure 42: Experiment of Mushroom dimension ordering using Dr (Rule Count) for dimensions and alphabetical order for categories.

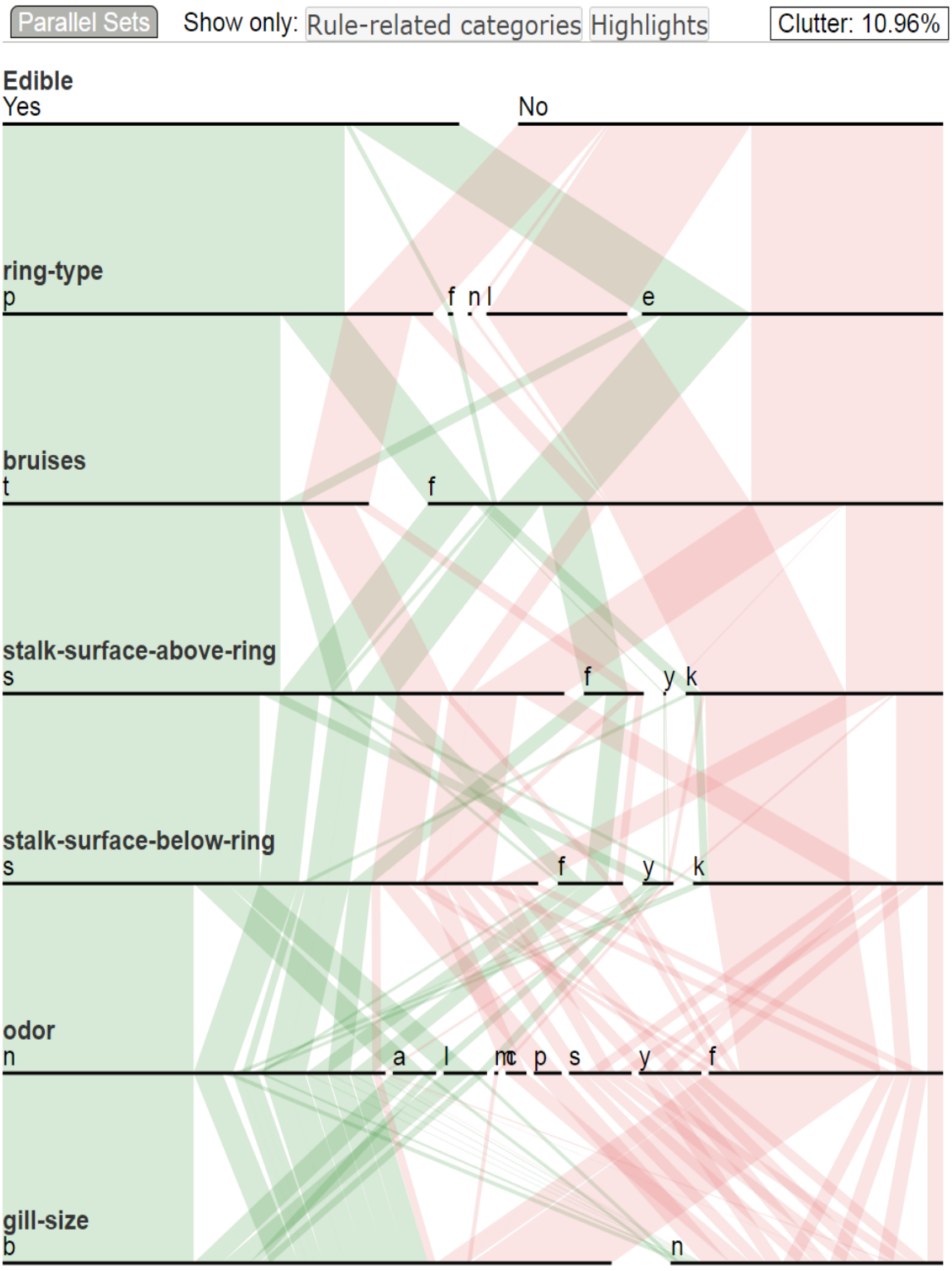

Figure 43: Experiment of Mushroom dimension ordering using Dr (Rule Count) for dimensions and Confidence for categories.

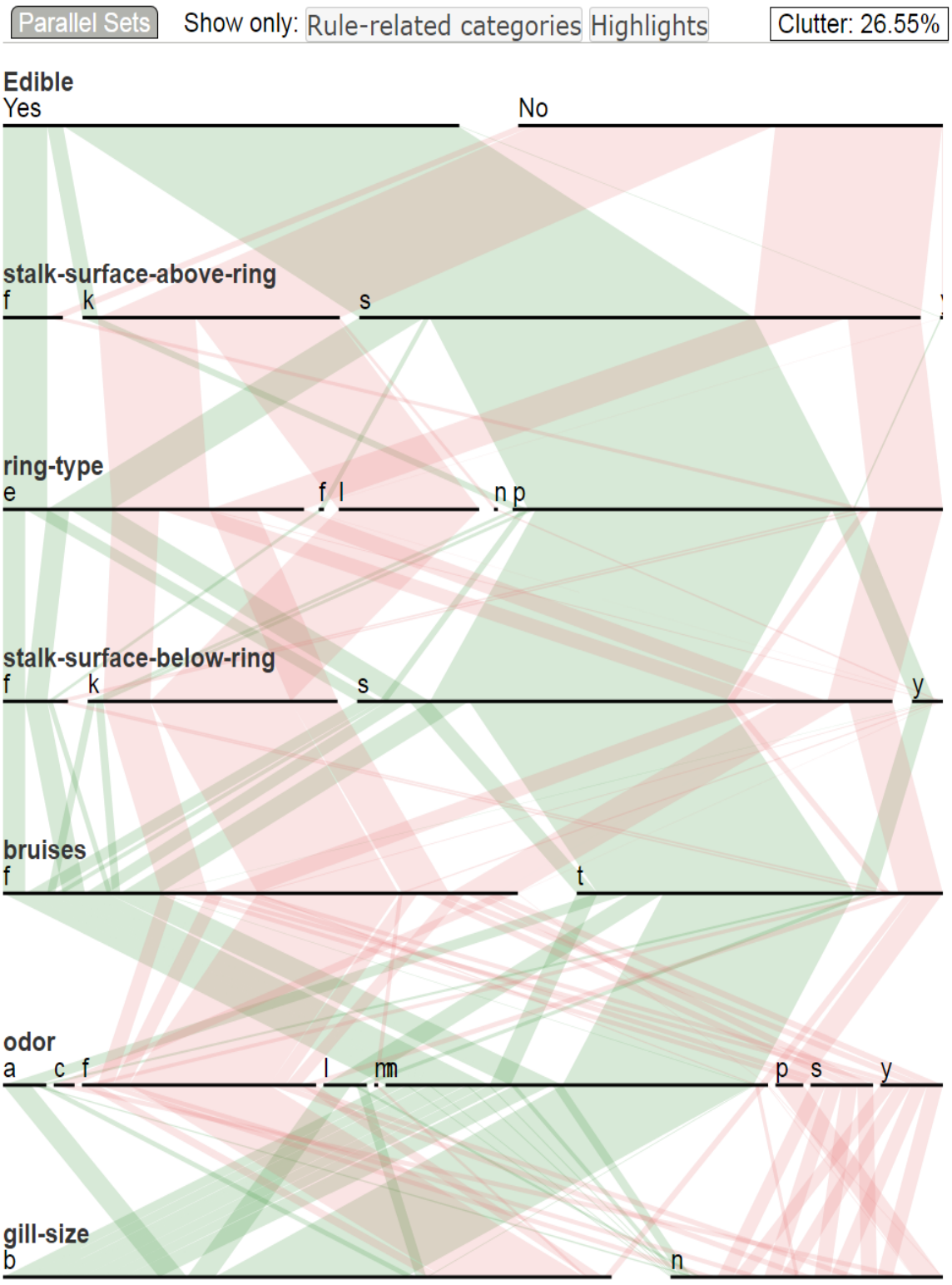

Figure 44: Experiment of Mushroom dimension ordering using Ds (Closeness) for dimensions and alphabetical order for categories.

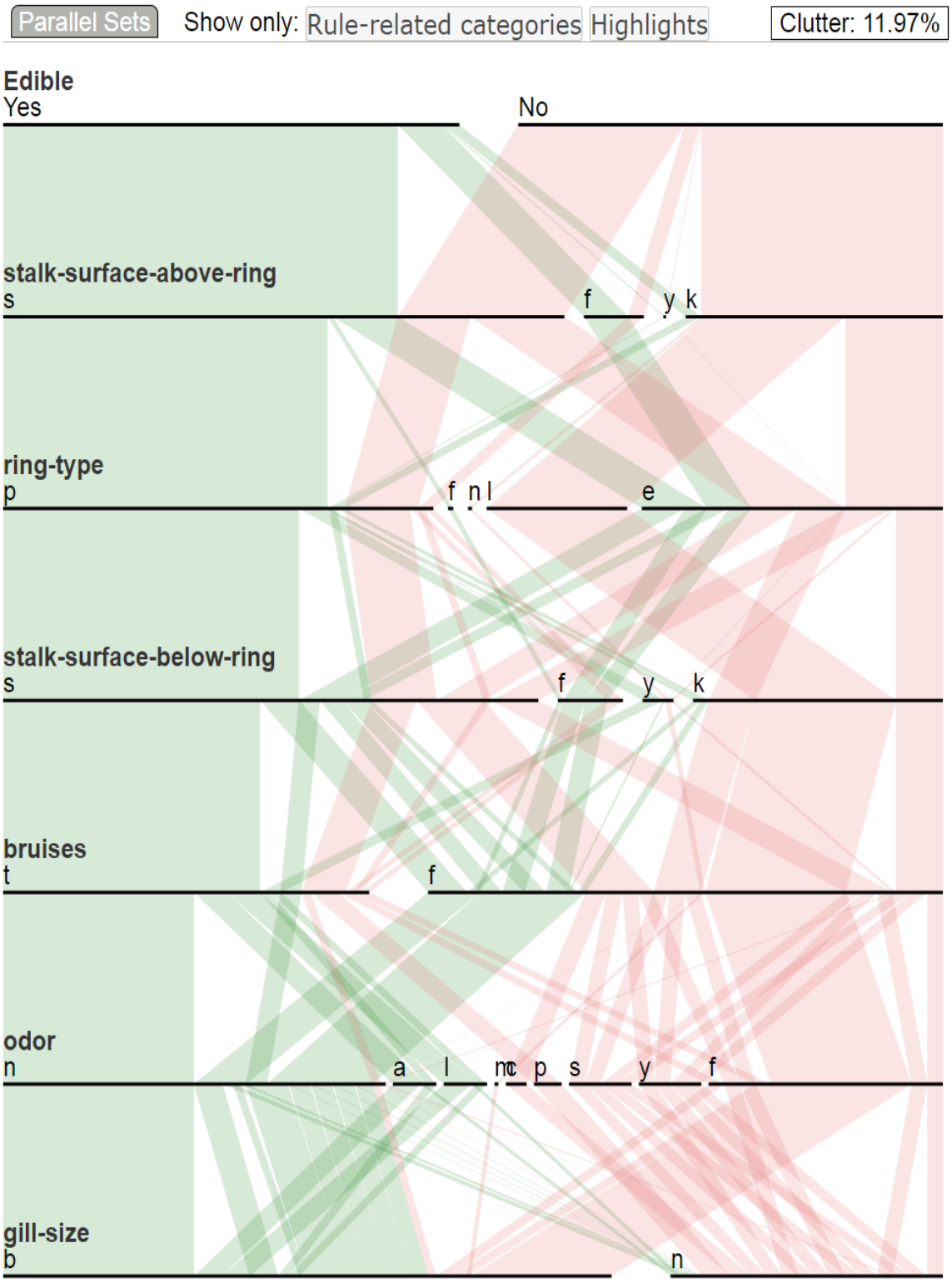

Figure 45: Experiment of Mushroom dimension ordering using Ds (Closeness) for dimensions and Confidence for categories.

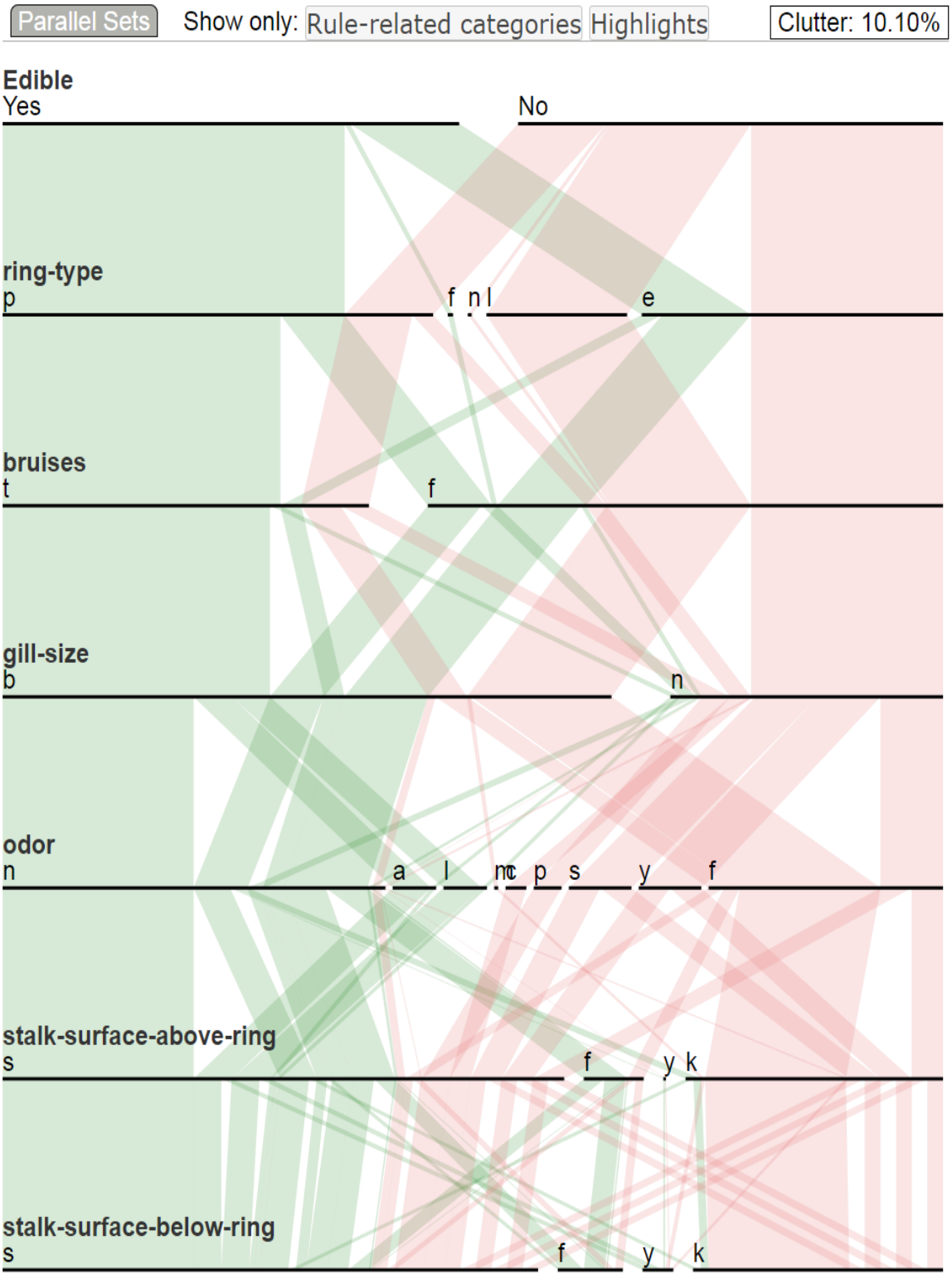

Figure 46: Experiment of Mushroom dimension ordering using YDc (Yes Cate. Count) for dimensions and Confidence for categories.

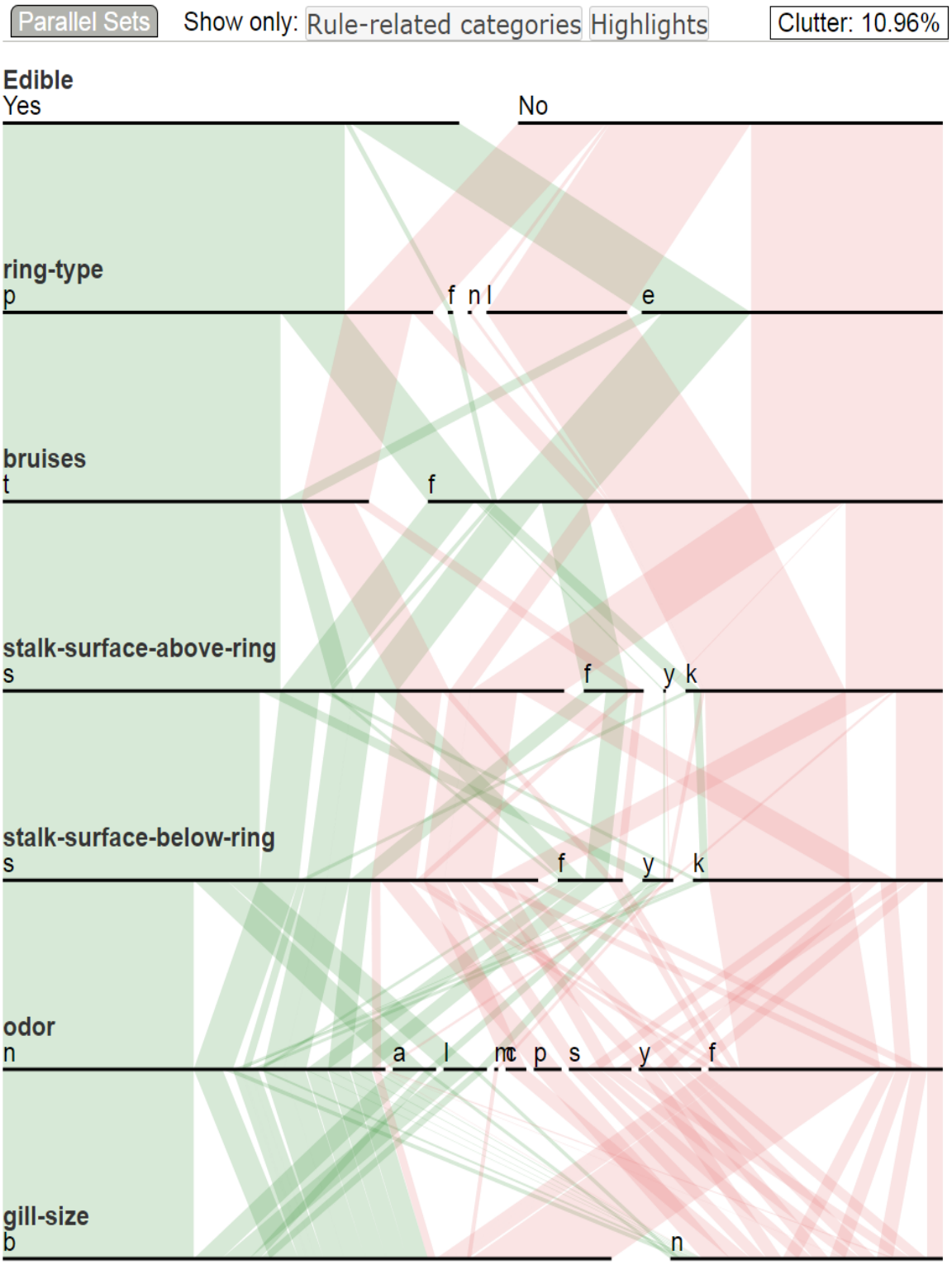

Figure 47: Experiment of Mushroom dimension ordering using YDr (Yes Rule Count) for dimensions and Confidence for categories.

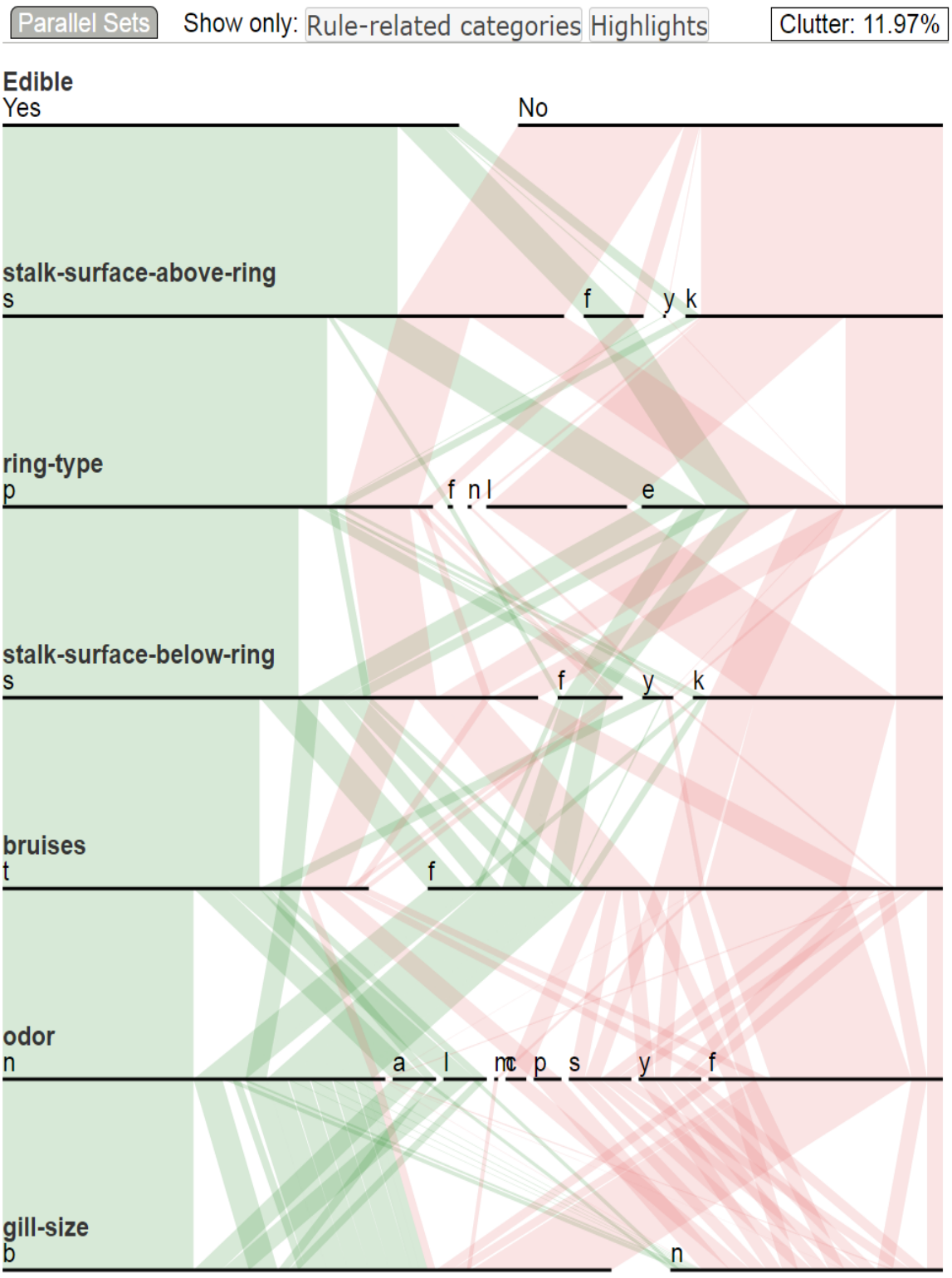

Figure 48: Experiment of Mushroom dimension ordering using YDs (Yes Closeness) for dimensions and Confidence for categories.

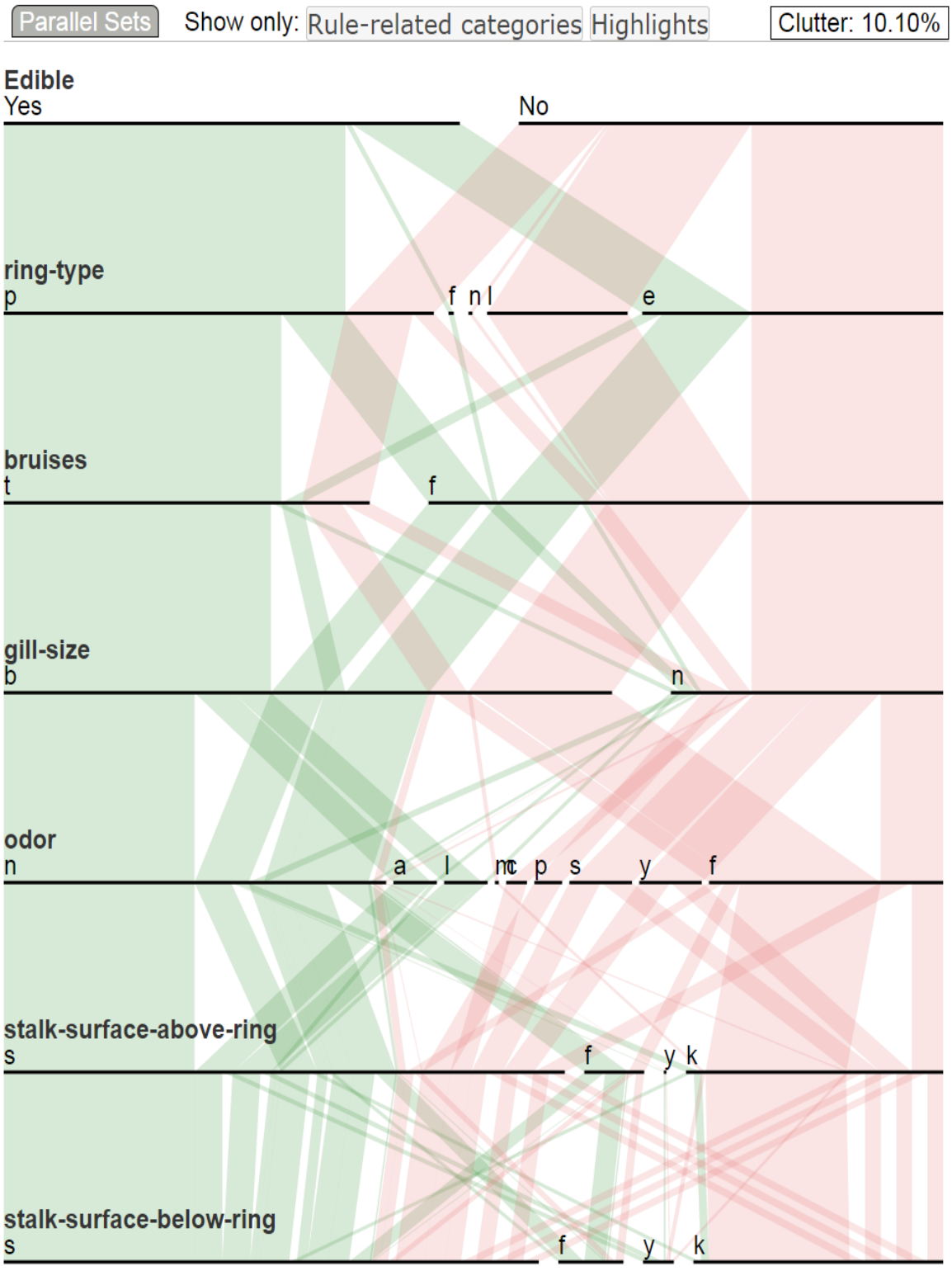

Figure 49: Experiment of Mushroom dimension ordering using NDc (No Cate. Count) for dimensions and Confidence for categories.

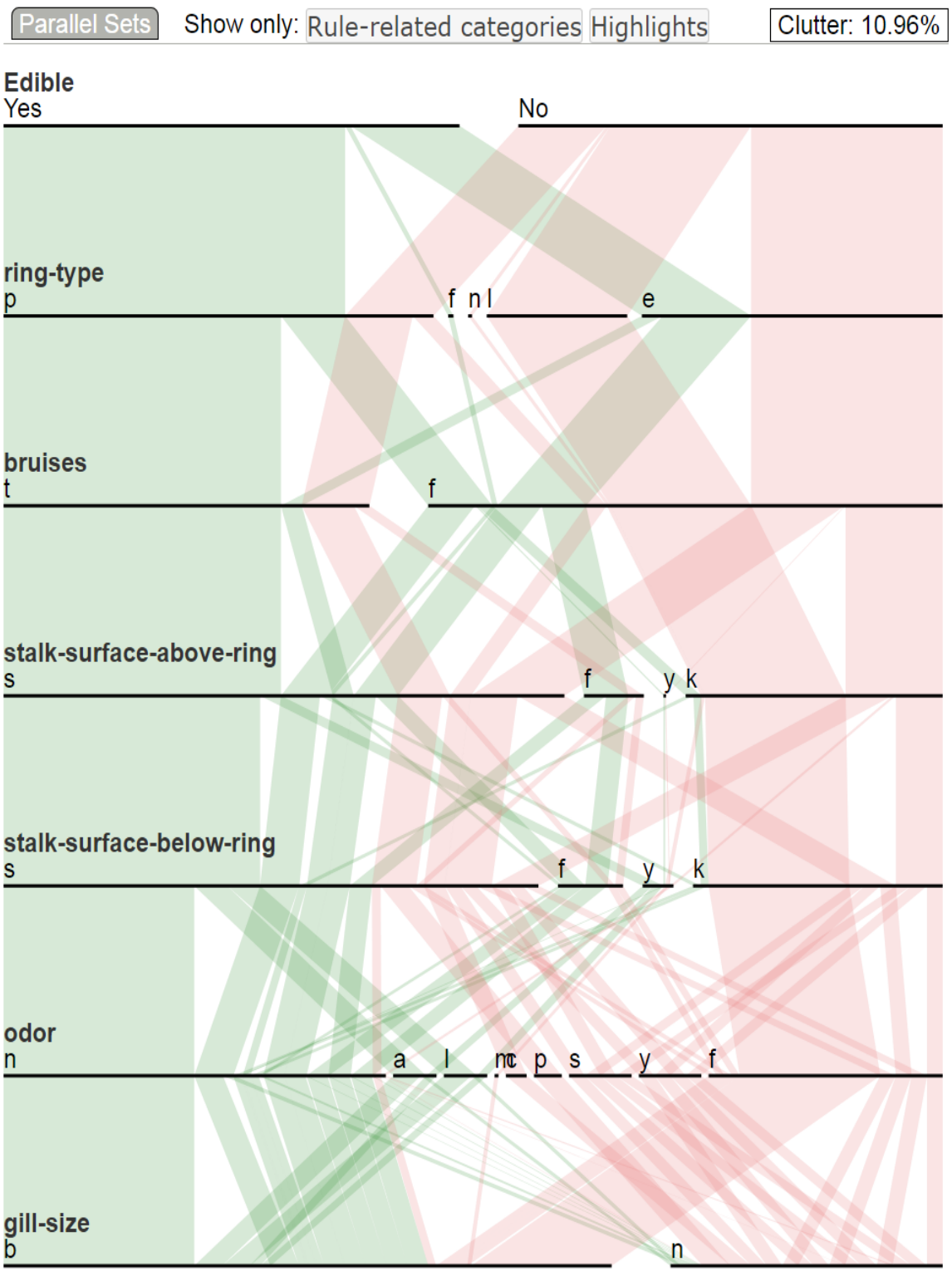

Figure 50: Experiment of Mushroom dimension ordering using NDr (No Rule Count) for dimensions and Confidence for categories.

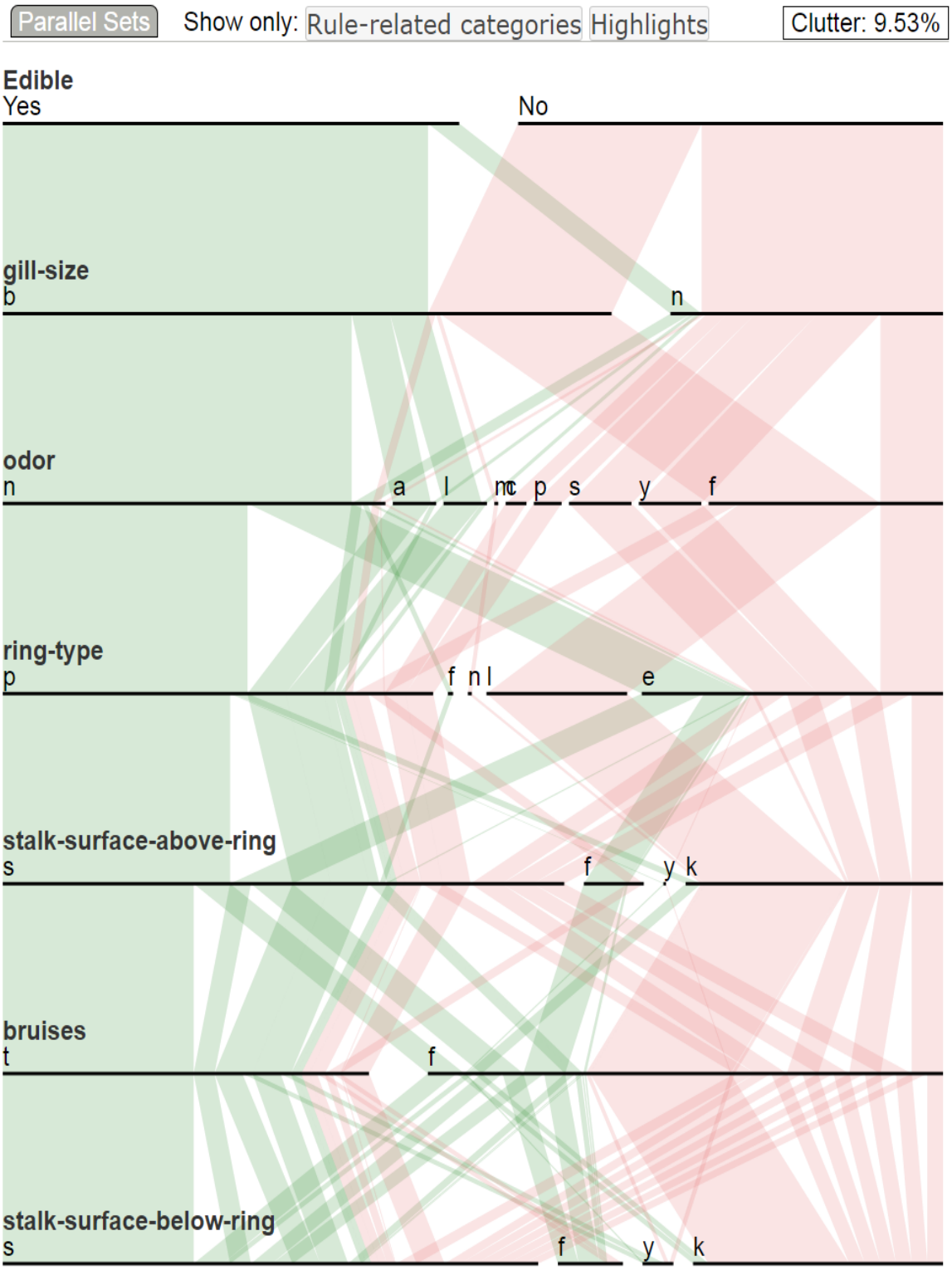

Figure 51: Experiment of Mushroom dimension ordering using NDs (No Closenss) for dimensions and Confidence for categories.

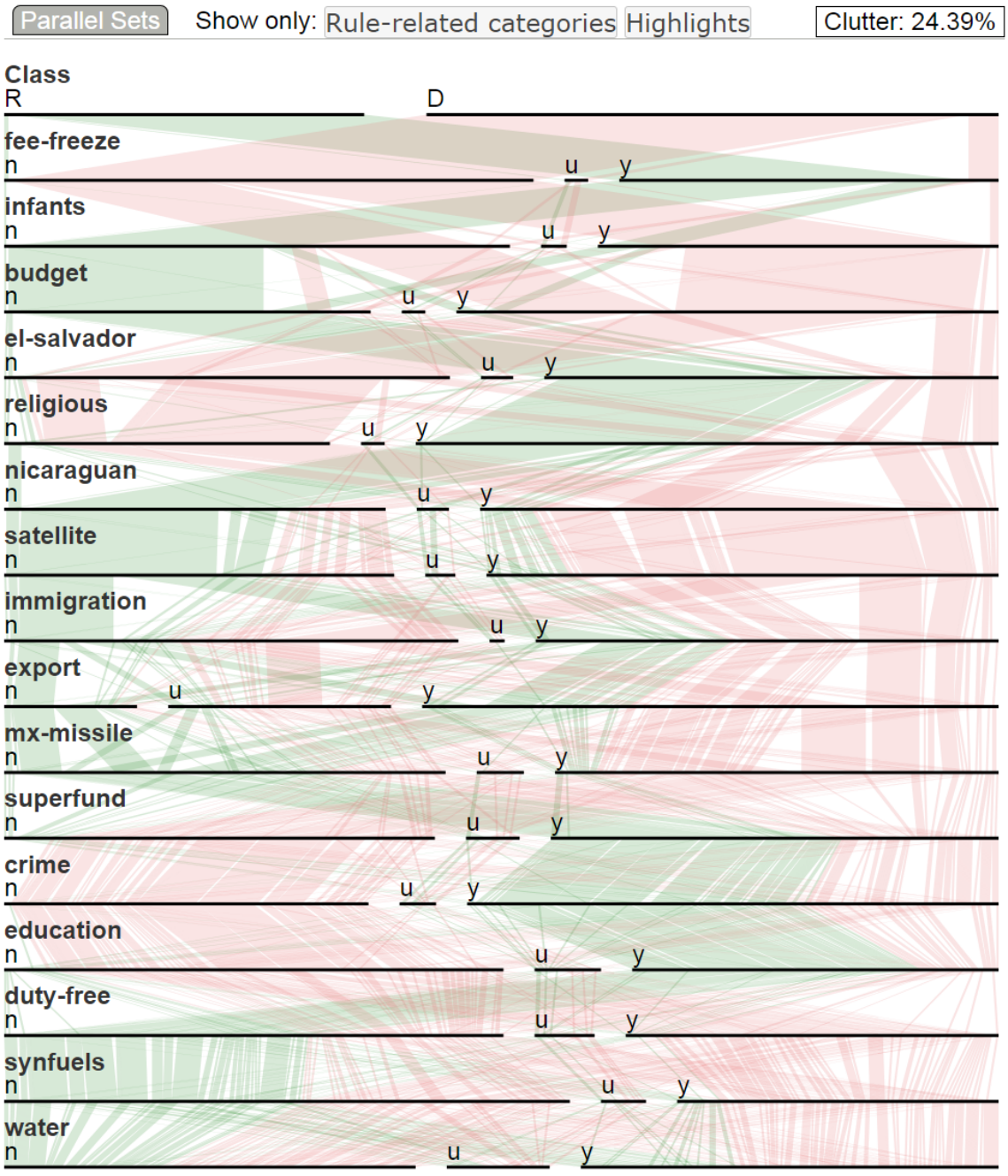

Figure 52: Experiment of Voting dimension ordering using Mutual Information for dimensions and alphabetical order for categories.

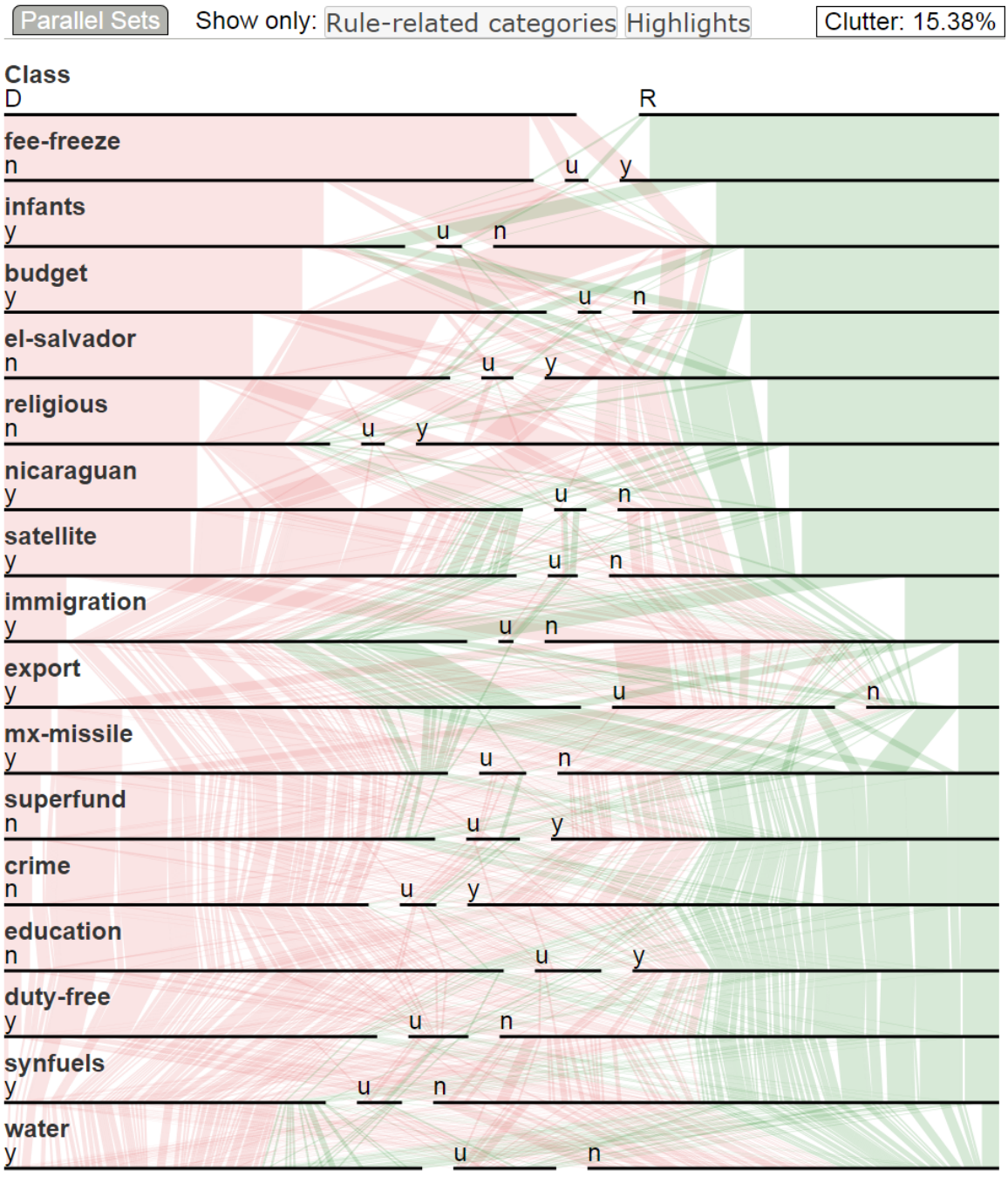

Figure 53: Experiment of Voting dimension ordering using Mutual Information for dimensions and Joint Entropy for categories.

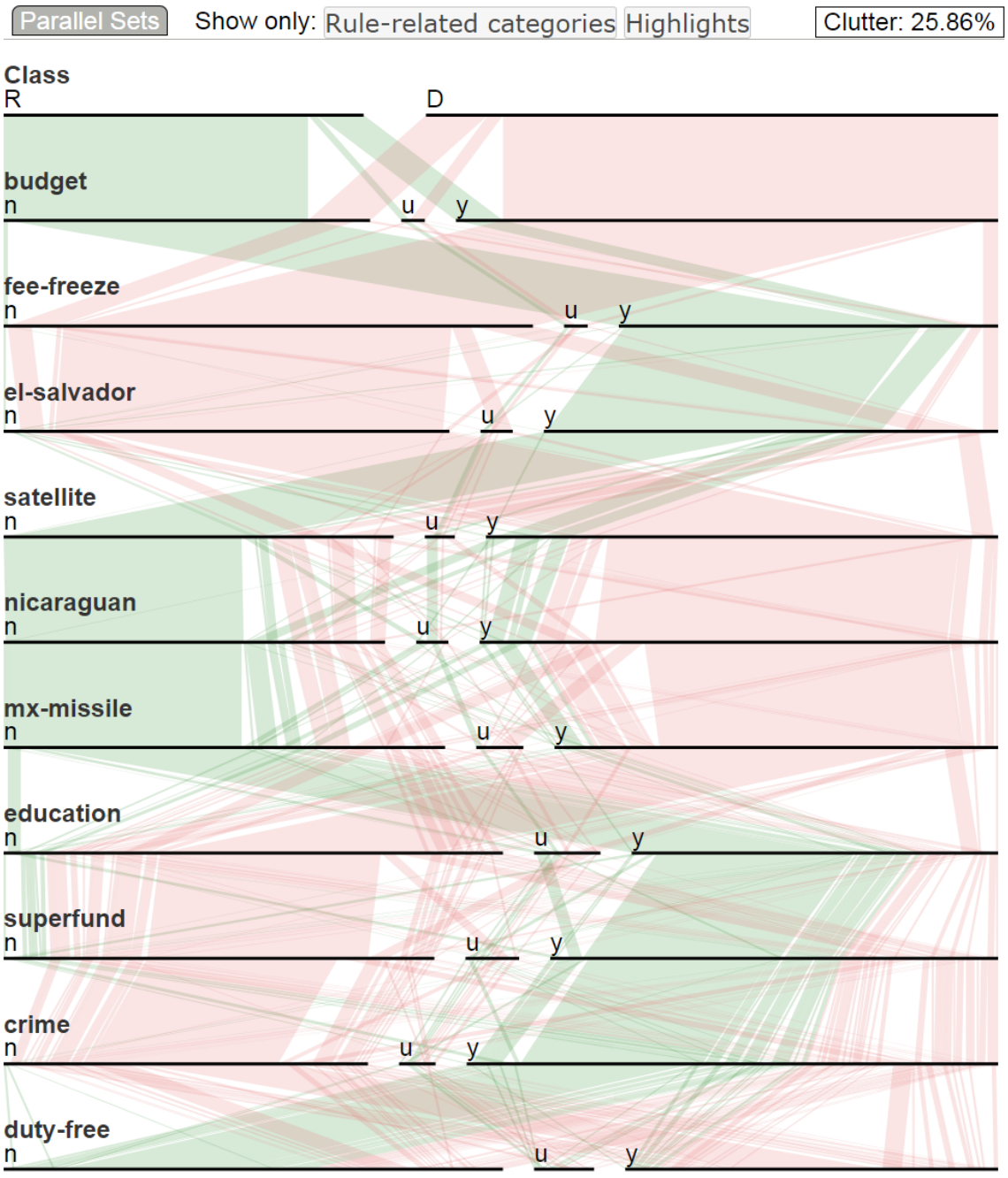

Figure 54: Experiment of Voting dimension ordering using Mutual Information for dimensions and alphabetical order for categories. Dimensions are reduced using association rules.

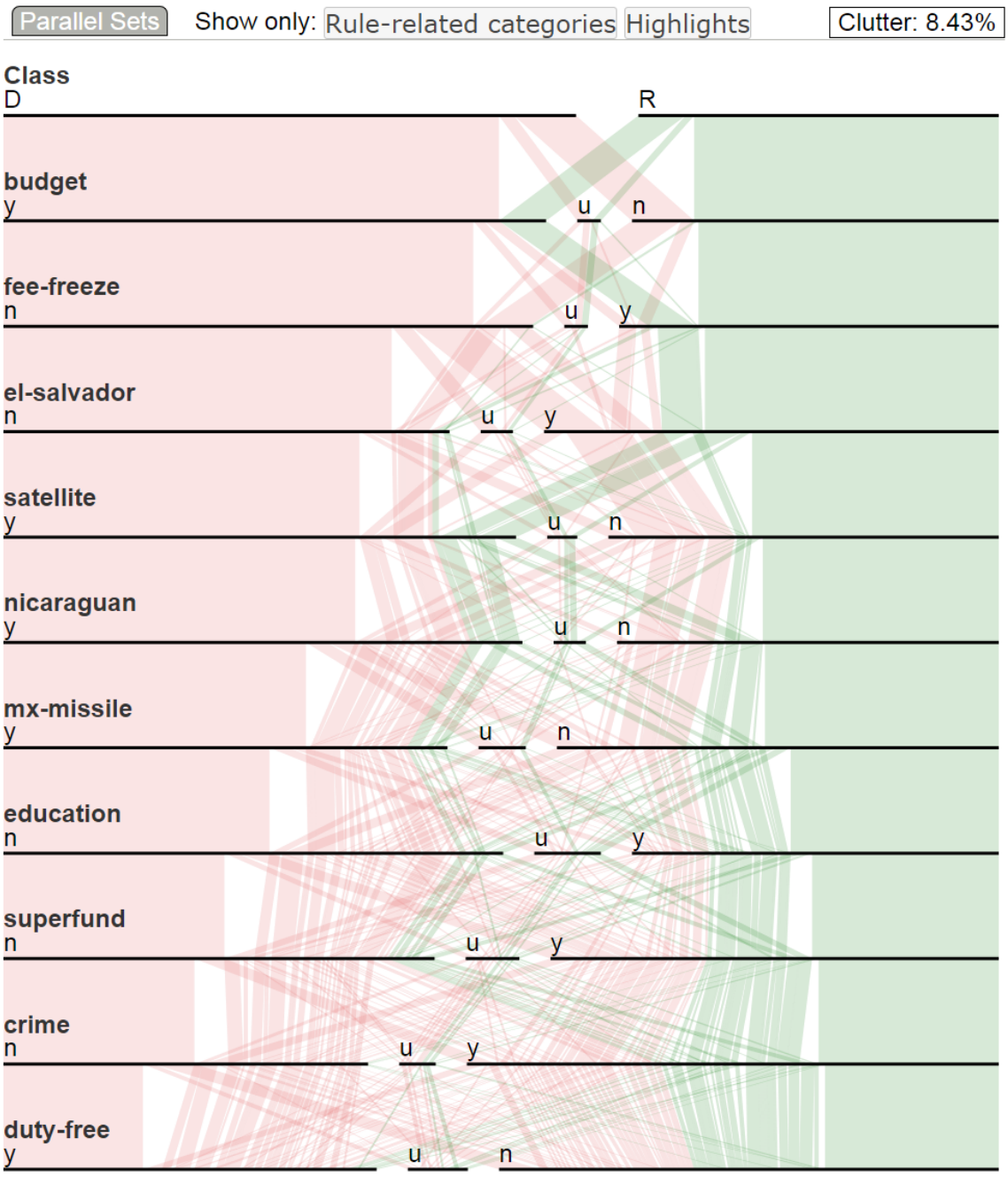

Figure 55: Experiment of Voting dimension ordering using Mutual Information for dimensions and Joint Entropy for categories. Dimensions are reduced using association rules.

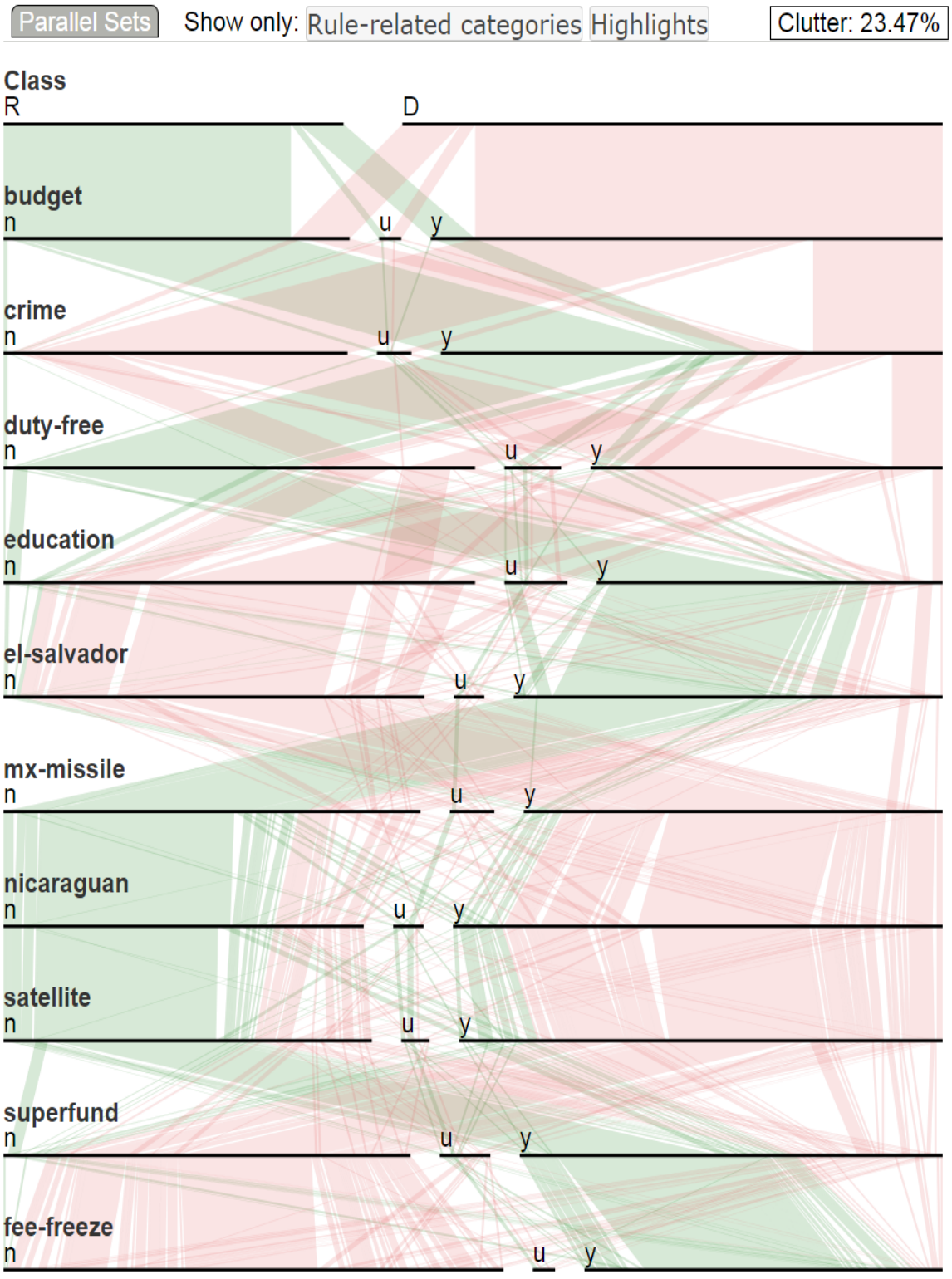

Figure 56: Experiment of Voting dimension ordering using Category Count for dimensions and alphabetical order for categories.

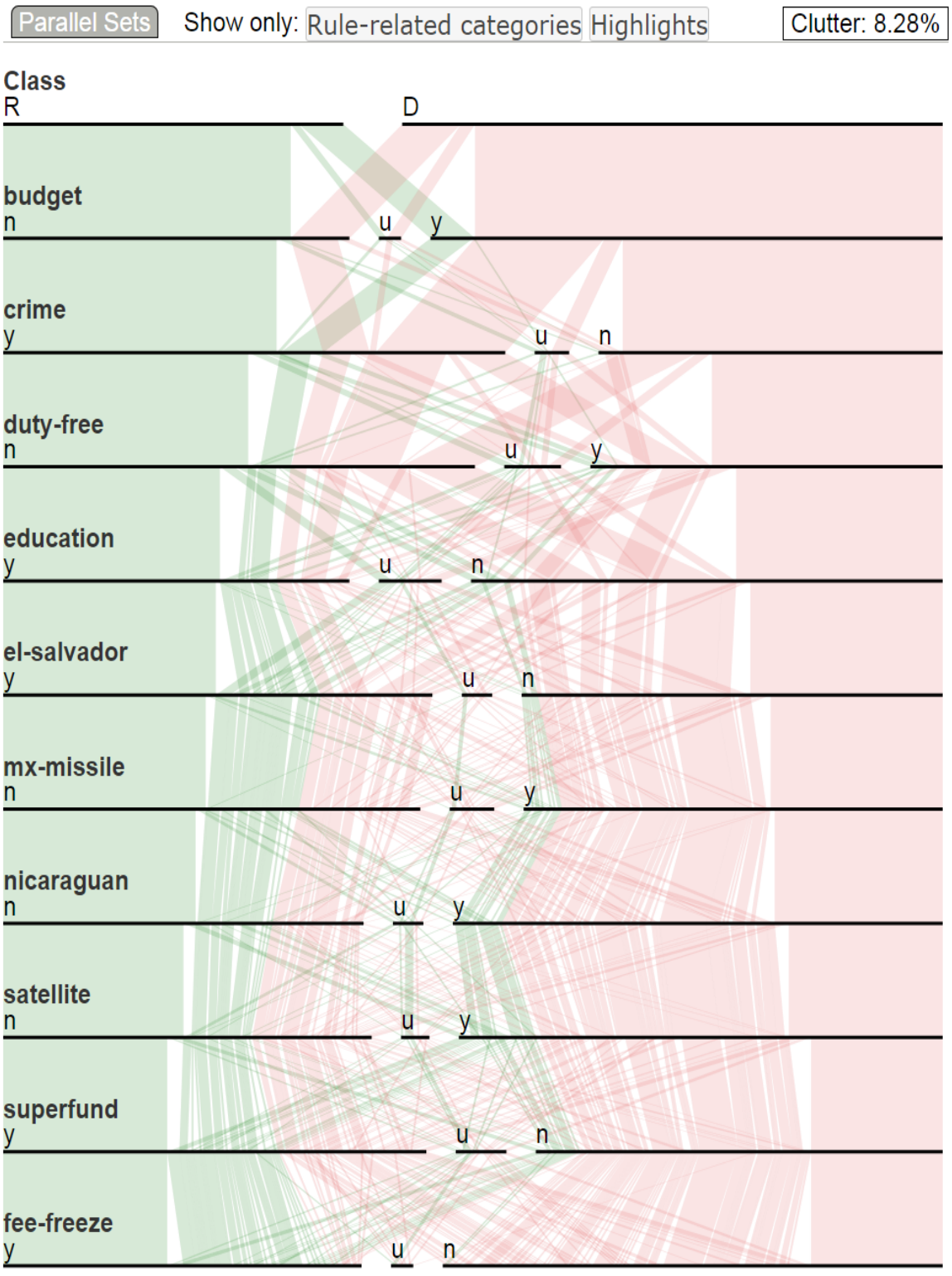

Figure 57: Experiment of Voting dimension ordering using Category Count for dimensions and Confidence for categories.

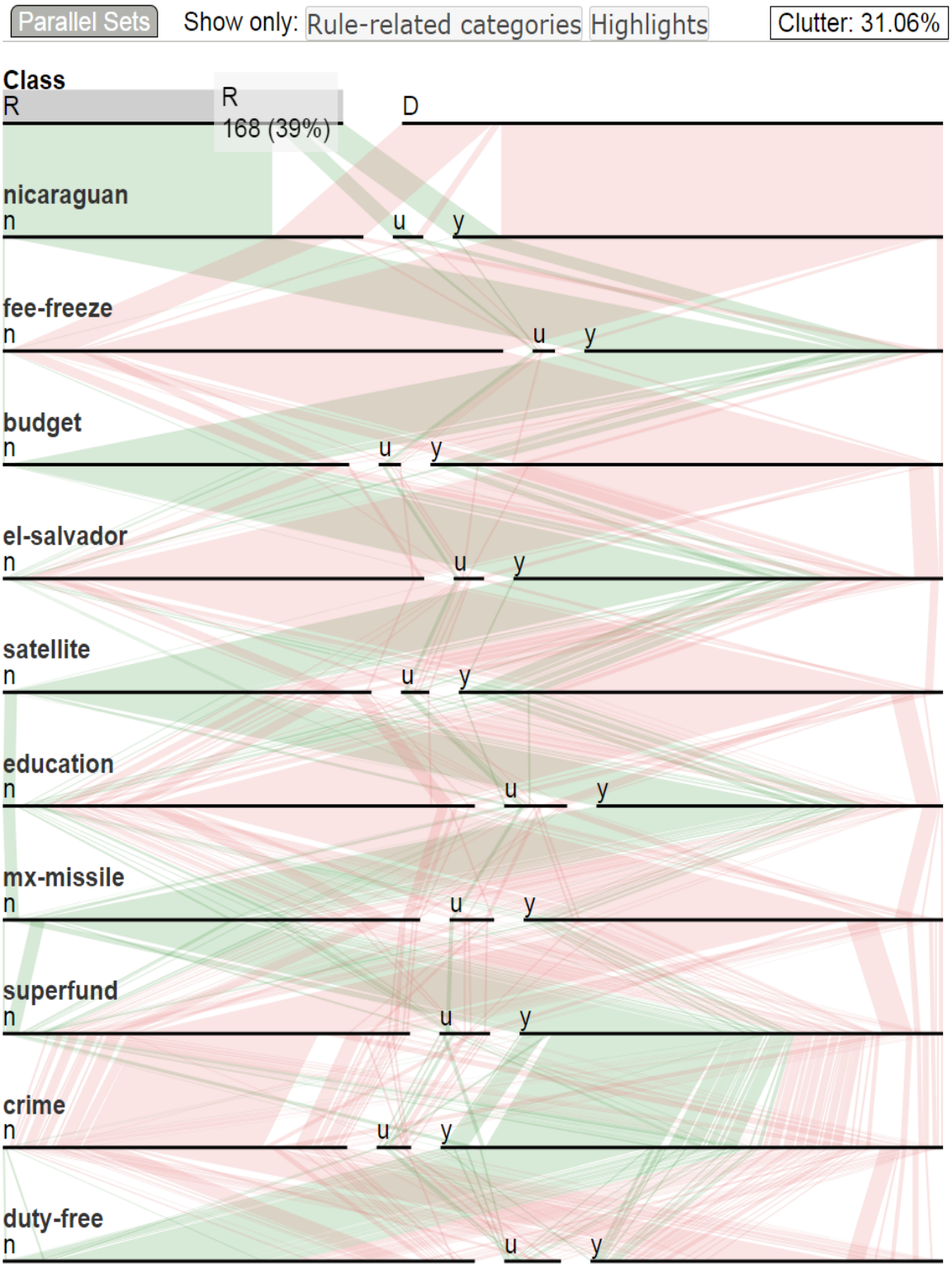

Figure 58: Experiment of Voting dimension ordering using Rule Count for dimensions and alphabetical order for categories.

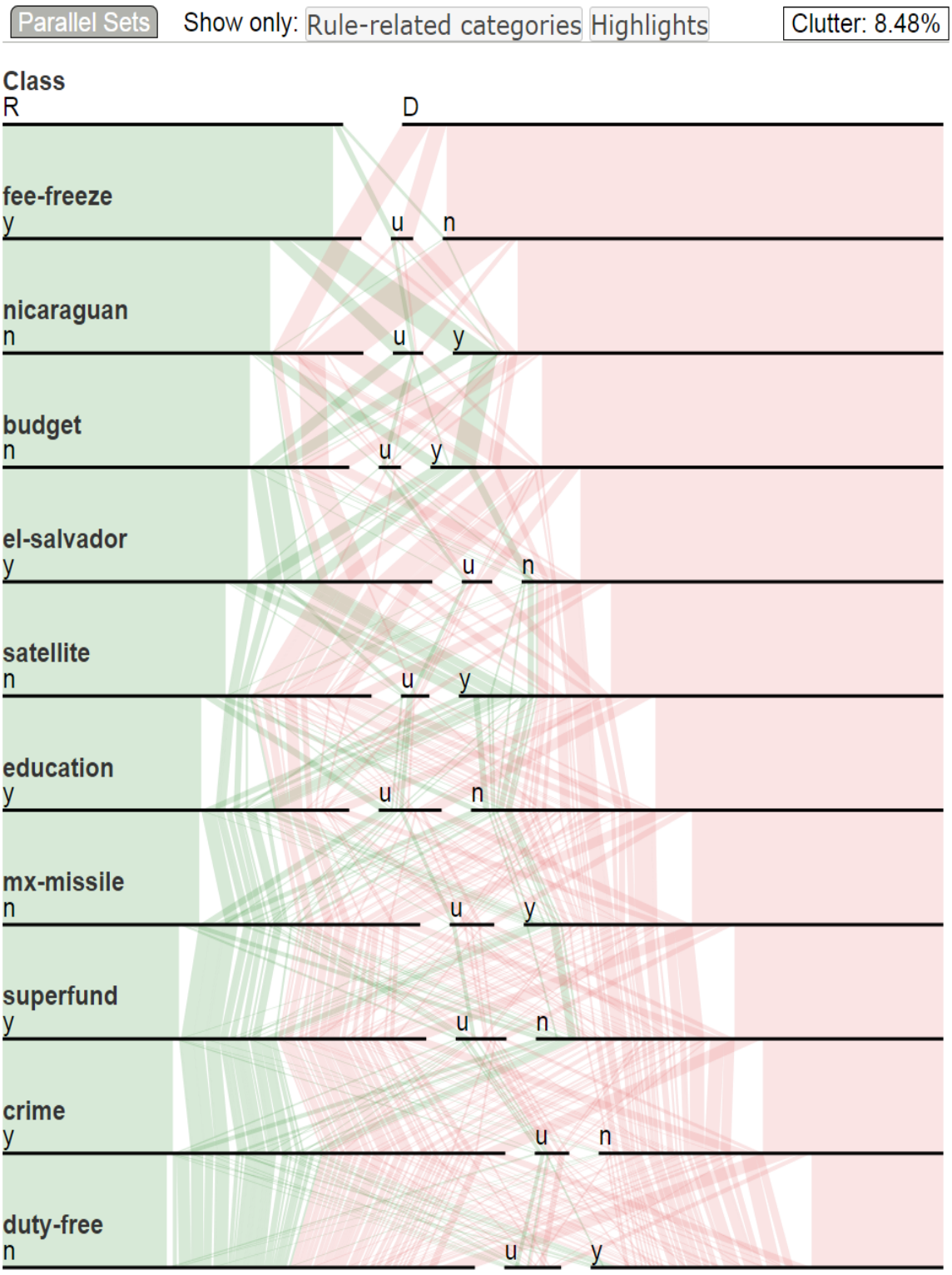

Figure 59: Experiment of Voting dimension ordering using Rule Count for dimensions and Confidence for categories.

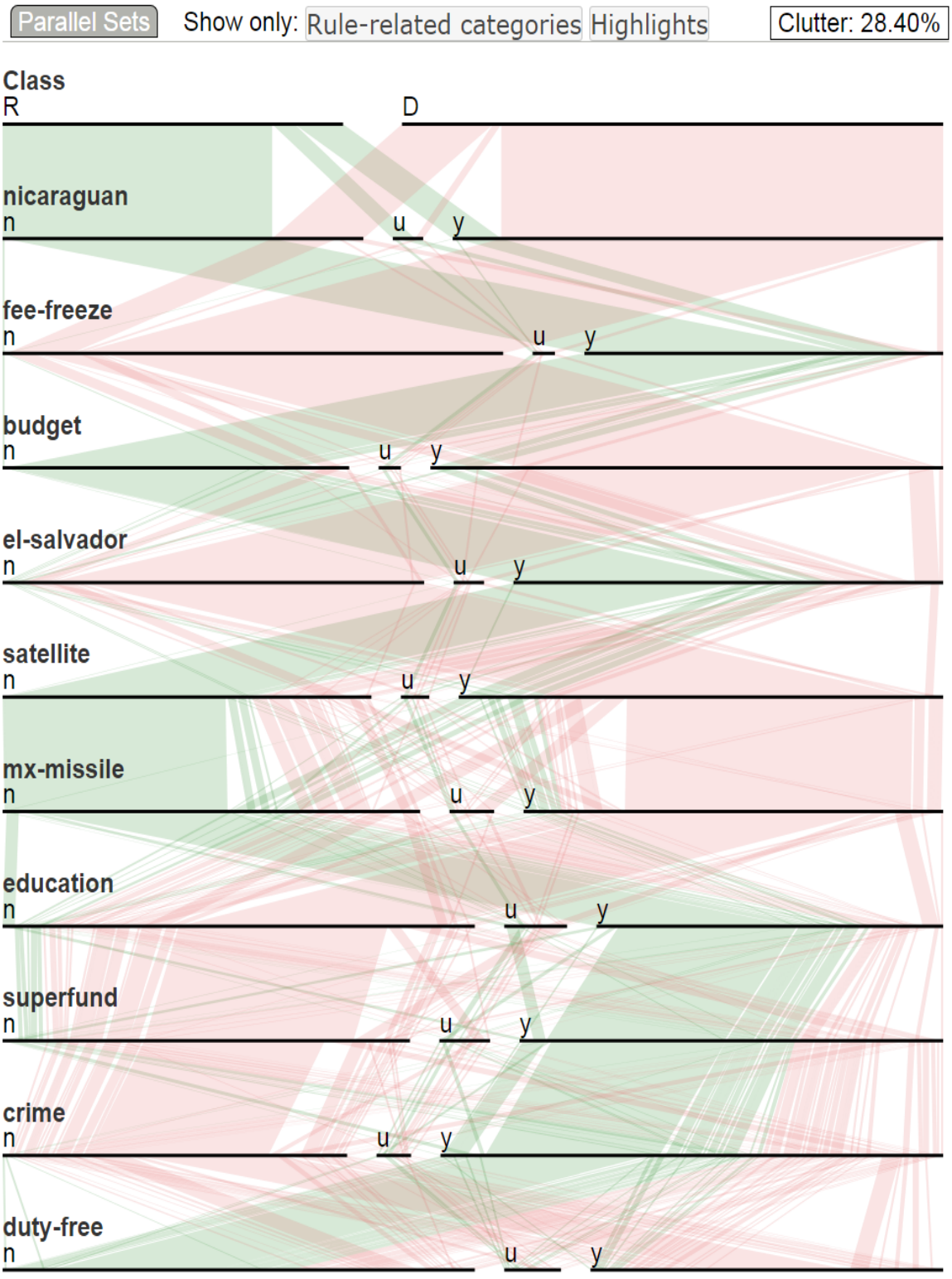

Figure 60: Experiment of Voting dimension ordering using Closeness for dimensions and alphabetical order for categories.

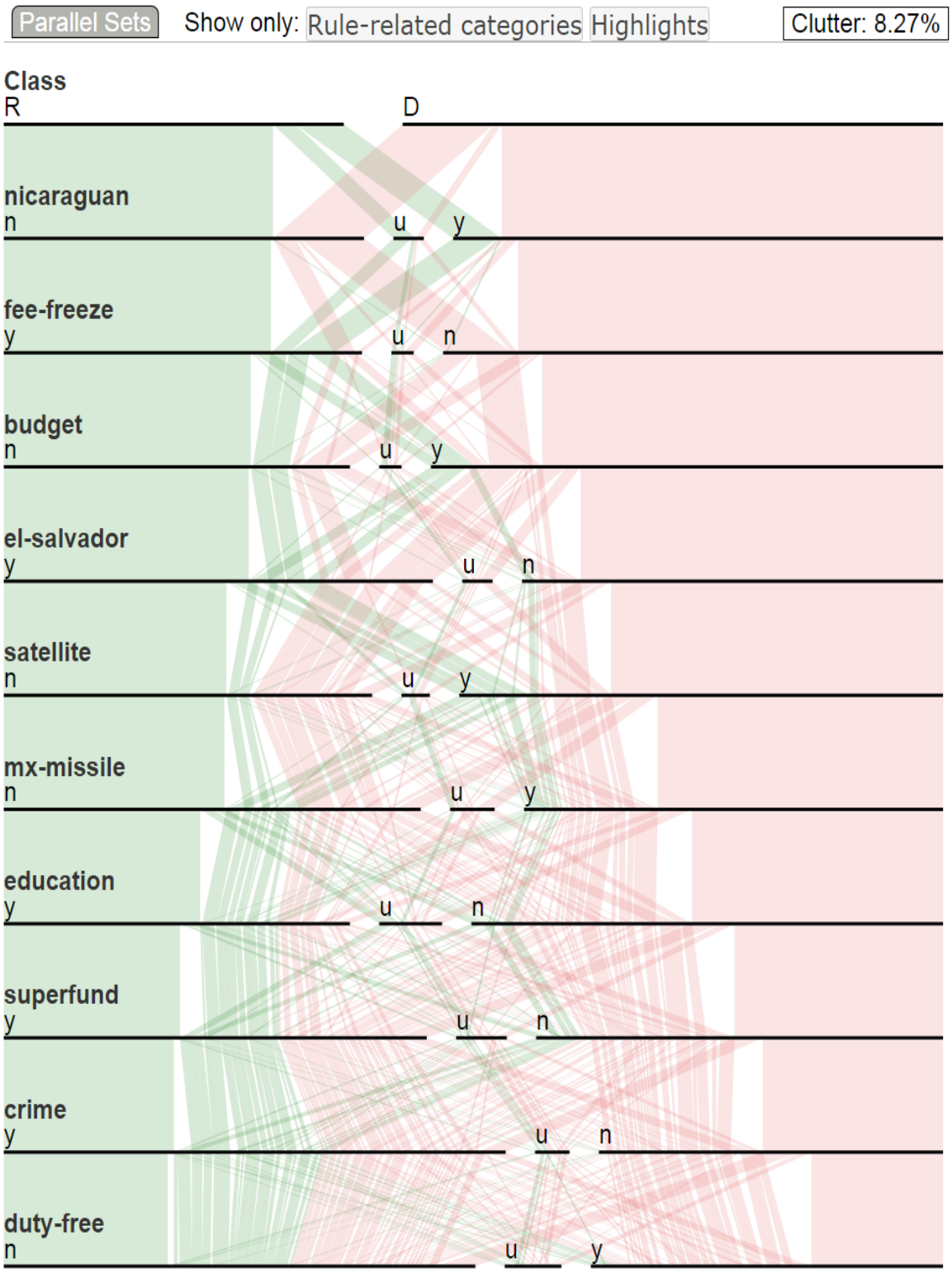

Figure 61: Experiment of Voting dimension ordering using Closeness for dimensions and Confidence for categories.

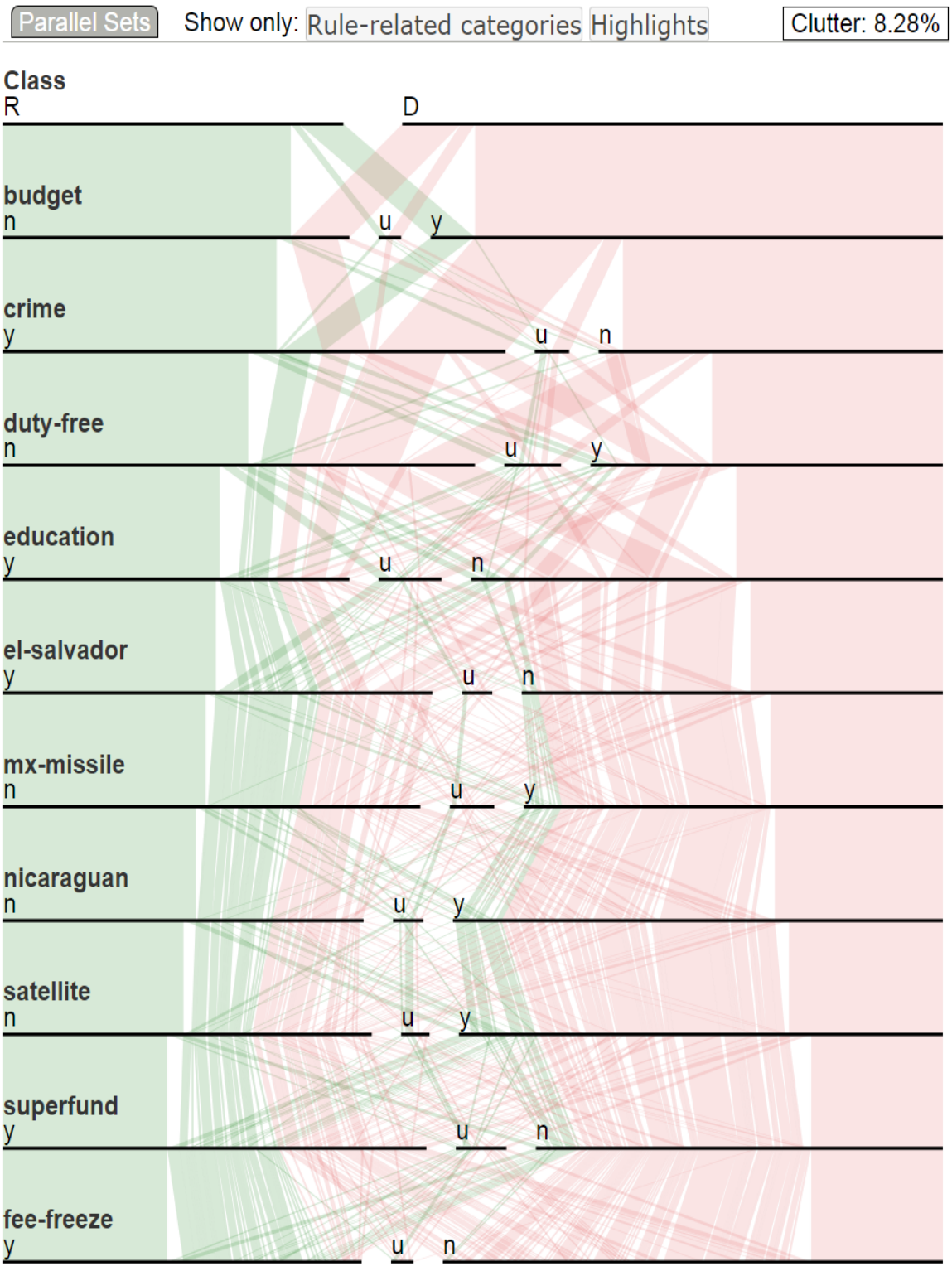

Figure 62: Experiment of Voting dimension ordering using Yes Cate. Count for dimensions and Confidence for categories.

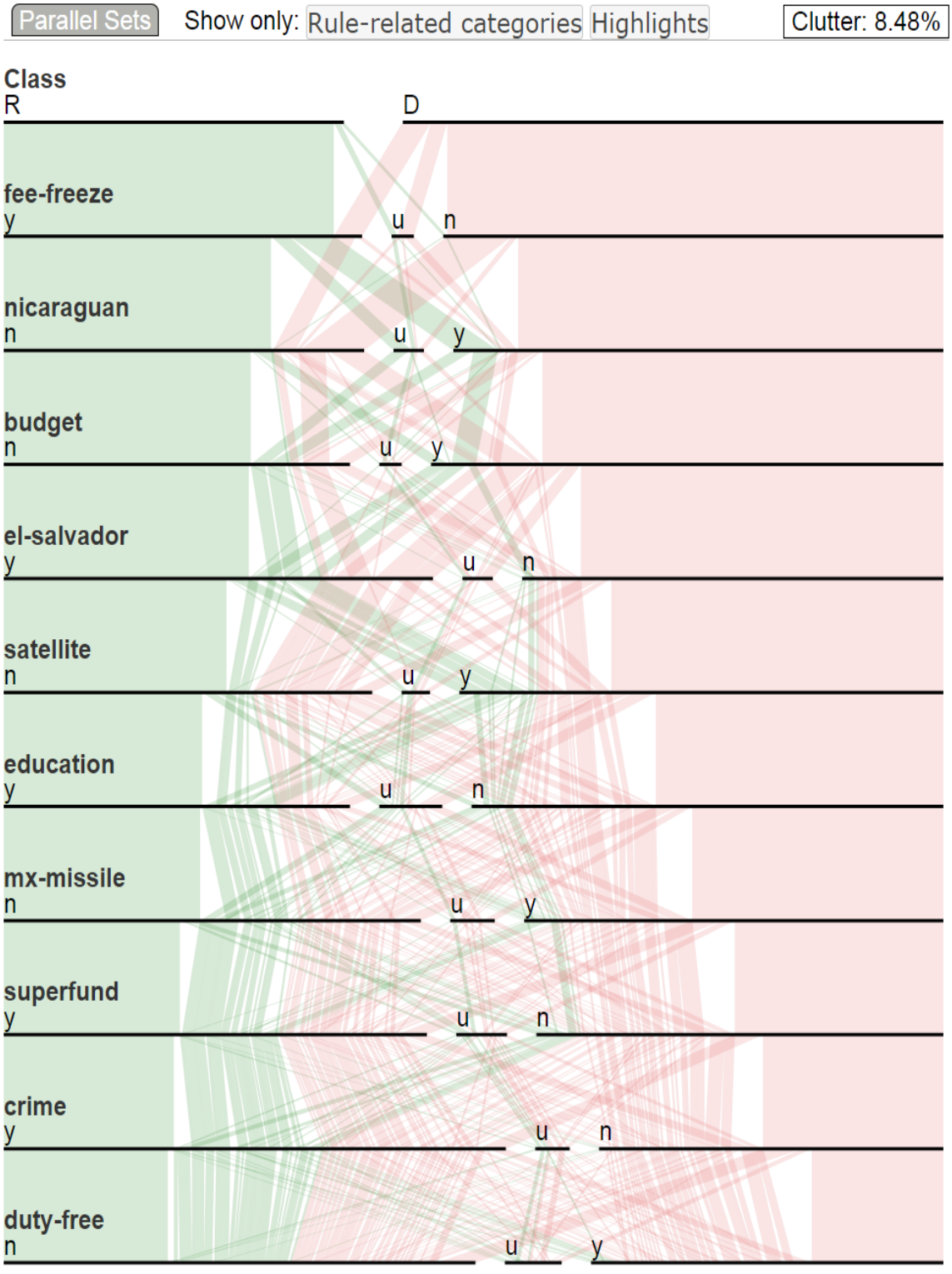

Figure 63: Experiment of Voting dimension ordering using Yes Rule Count for dimensions and Confidence for categories.

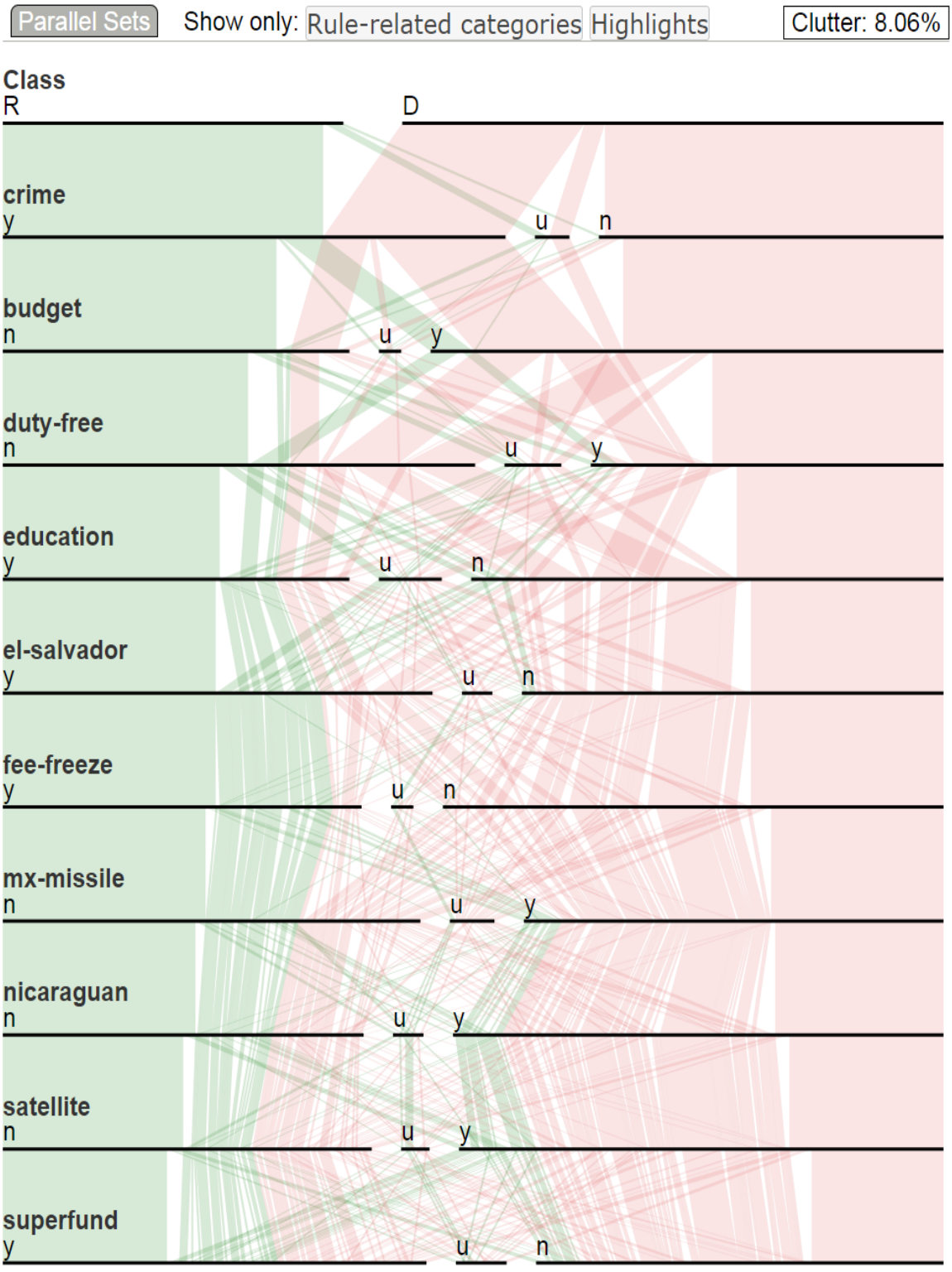

Figure 64: Experiment of Voting dimension ordering using Yes Closeness for dimensions and Confidence for categories.

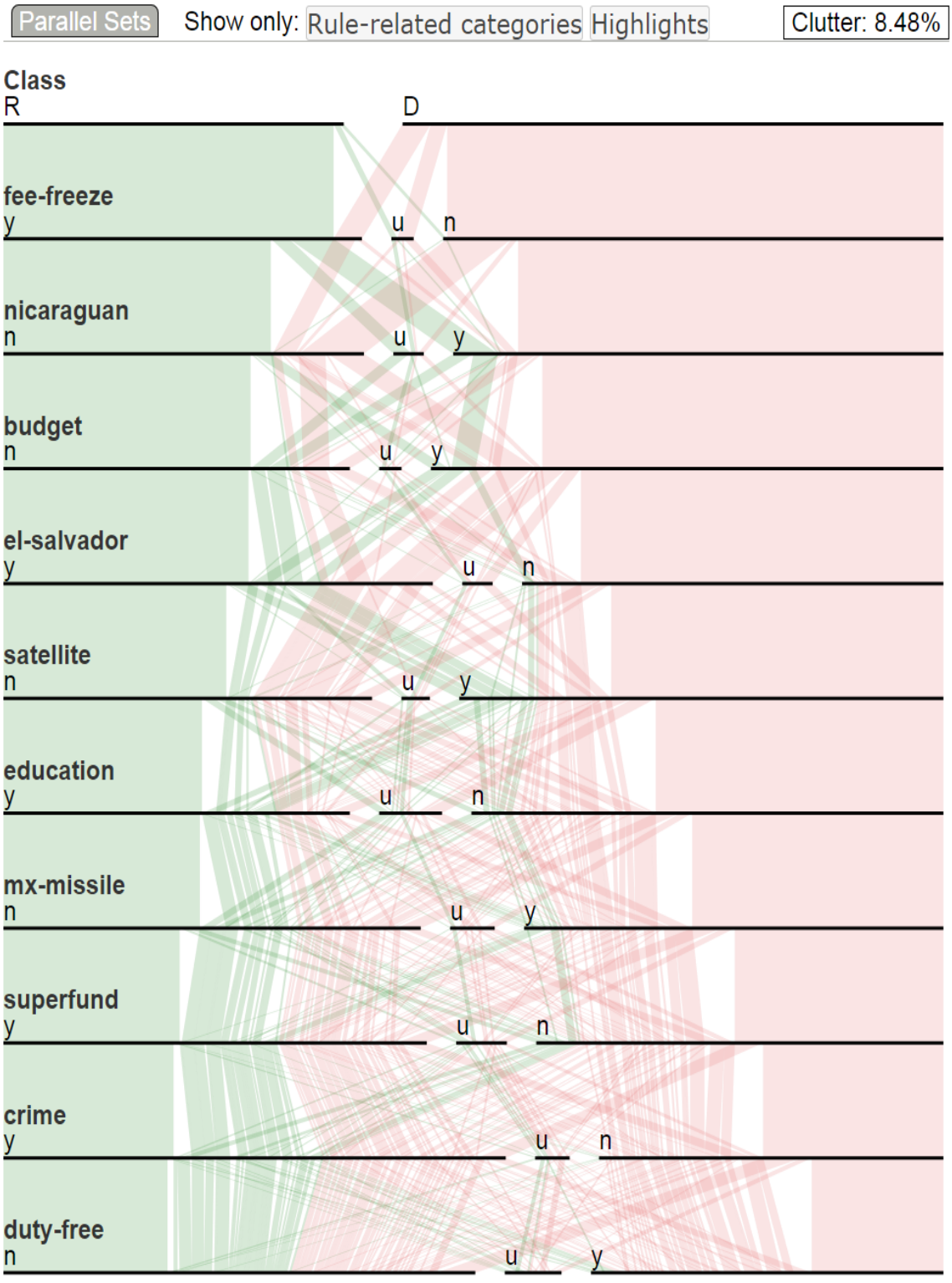

Figure 65: Experiment of Voting dimension ordering using No Cate. Count for dimensions and Confidence for categories.

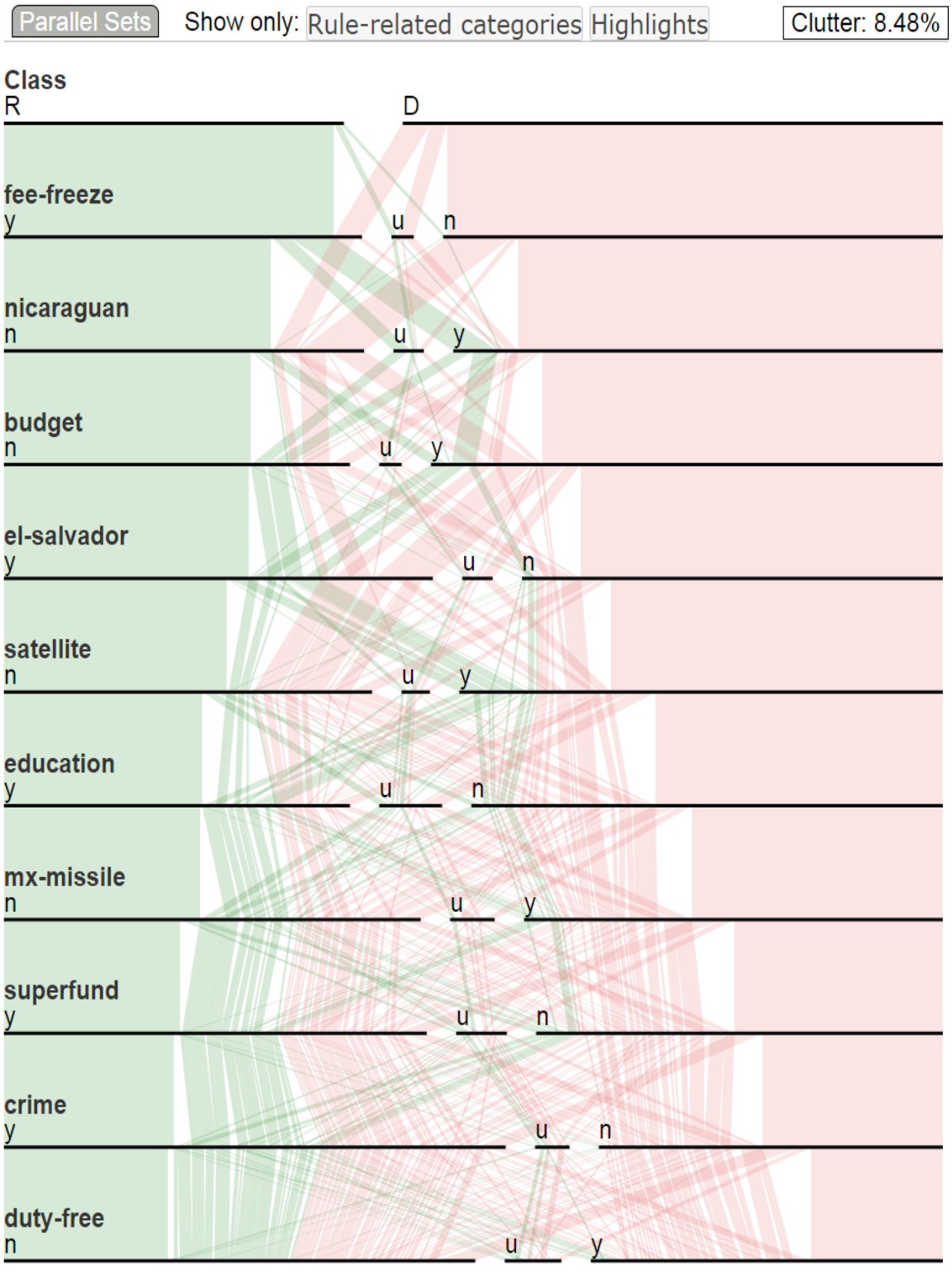

Figure 66: Experiment of Voting dimension ordering using No Rule Count for dimensions and Confidence for categories.

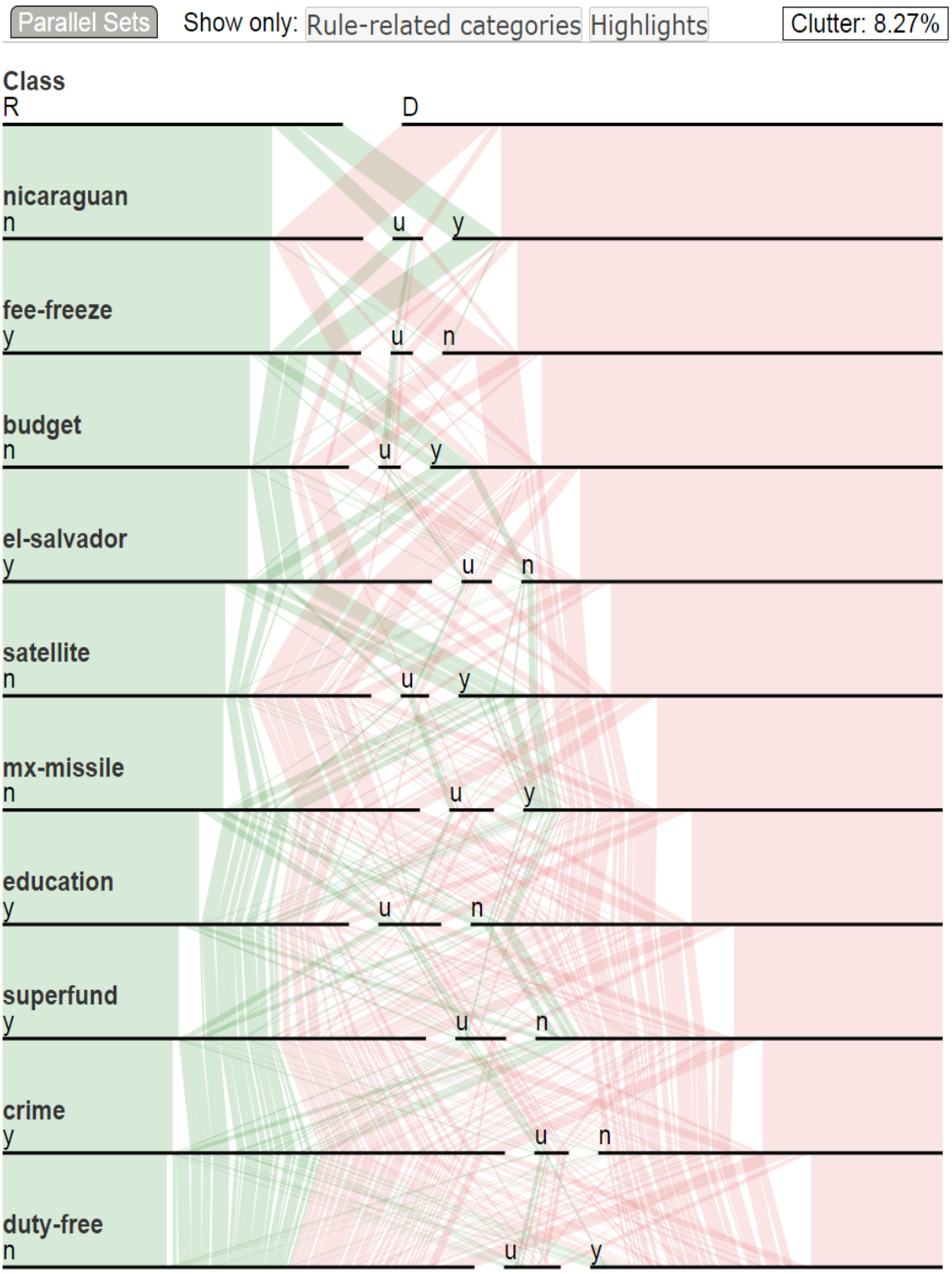

Figure 67: Experiment of Voting dimension ordering using No Closeness for dimensions and Confidence for categories.

4.16 Discussion and Conclusion

The aim of this project was to investigate parallel sets and consider methods of overcoming the problem of clutter when analyzing high-dimensional categorical data with dichotomous outcome. In developing the clutter reduction technique, it has become apparent that many factors must be considered in terms of user cognition and interpretation. We need to balance the aesthetic pleasingness and semantic representations of parallel sets. Heuristic approach might result in an optimal order of dimensions and categories that the clutter is reduced at the maximal level. However, the semantic groups of dimensions and categories might be separated and across around, which will increase the difficulty of interpretation of the relationship.

The clutter metric could be improved for a more precise measure. We did not consider the perception effect and cognition load of ribbon width with regard to the pattern discovery in Parallel Sets.

In this chapter, we present a novel dimension and category ordering method using association rule for Parallel Sets of a high-dimensional categorical data with dichotomous outcome. It integrates an association rule mining for association extraction. It enhances the exploration with three broads of ordering methods and interactions that allow users to easily find the relationship using Parallel Sets, and allows users to effectively and efficiently examine interesting association between the categorical variables.

CHAPTER 5: CONCLUSION

High-dimensional data visualization may well be one of the most researched topic of Visual Analytics. The primary goals of it are to explore and/or explain the data. Dimension reduction is used often to support the goals. However, with a dichotomous outcome dimension in the data, new challenges have arisen:

- Not all observed variables are important for understanding the outcome of interest. Analysis tools need to help explain how the outcome variable changes when any one of the explanatory variables is varied.
- High-dimensional regression modeling seeks the most parsimonious model that still explains the data. Challenges such as overfitting, multicollinearity, and confounding are experienced with the high dimensionality.
- Visual analytics approaches need to facilitates the model building.
- The association between categorical dimensions needs an effective exploration.

In this dissertation, all these points were explored: We proposed interactive visual analytics approaches, combined with statistical and data mining algorithms to make analysis of complex inter-correlation understandable, without the help of scripting. We demonstrated how visual design can improve the understanding in terms of the steps of high-dimensional regression model building. We explained how these ex-

planatory variables are related and affect the outcome variable by employing various visualizing techniques and rich interactions. We also demonstrated how to take advantage of association rule mining for categorical data exploration enhancement in Parallel Sets.

We have presented three visual analytics systems for high-dimensional data with dichotomous outcome. Two of them are focused on continuous data regression modeling, the other on categorical data exploration analysis. We discuss these as design studies, with a detailed analysis regarding the tasks and problems.

We discussed the requirements and considerations of global task and local task for EPA research project, including how to efficient examine potential correlations between the hundreds of chemicals and certain types of birth defects, and how to explore the vulnerable groups of people and their risk factors using Rose Graph and Dual Axes Parallel Coordinates.

As the investigation going deeper, we demonstrated that how visualization improves the understanding of statistical procedures and help facilitate regression model building. We discussed how to visually build a stable model for heterogeneous data where chemicals were treated as continuous variables and demographic and social-economic information as categorical variables. We elaborated on how the visual analytic approach distinguishes true risk factors from confounders, how these variables work together to influence the outcome variable. We also explored the effect modification of these categorical variables on the relationships with the outcome variable. We conducted our case studies with a senior Ph.D. student and an academic professor of statistics from a different institution. The results of the case studies affirm our claim

that the visual analytics approach have a clear benefit in the stable model building. The professor also gave a very positive feedback.

We also explored the high-dimensional categorical data exploration. A new dimension reordering method was introduced to reduce clutter for one of the modern visualization, Parallel Sets, with the help of association rule mining. We discussed the drawbacks of existing visualizing techniques. We chose to improve the legibility of the representation of Parallel Sets when a high-dimensional categorical data involved. Association rule mining brings such associated dimensions and categories together that can be used to create a new order of dimensions for Parallel Sets. Quantitative metric has indicated that the new order significantly reduced the clutter in Parallel Sets for high-dimensional categorical data exploration. We have balanced the aesthetic and semantic representations and conducted case studies on two benchmark datasets. We also evaluated our approach comparing the entropy-based method recently used in the visualization community. The studies indicated that our proposed method has improved the clutter reduction better while keeping the associated dimensions and categories together.

Looking forward, we would like to explore the benefits of visual analytics in probabilistic graphical models over a high-dimensional space. Graphs are an intuitive way of visualizing the relationships between many variables with nodes corresponding to variables and edges representing statistical dependencies between the variables. We would also like to analyze variable relationships in the context of geospatial. By applying visual analytic to spatial statistics and spatial regression modeling, the relationship finding could become more interesting.

REFERENCES

[1] H. Abdi. Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, pages 792–795, 2003.

[2] A. Agresti. *An introduction to categorical data analysis*, volume 423. Wiley-Interscience, 2007.

[3] A. Agresti and M. Kateri. *Categorical data analysis*. Springer, 2011.

[4] J. Alsakran, X. Huang, Y. Zhao, J. Yang, and K. Fast. Using entropy-related measures in categorical data visualization. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific*, pages 81–88. IEEE, 2014.

[5] B. Alsallakh, W. Aigner, S. Miksch, and M. E. Gröller. Reinventing the contingency wheel: Scalable visual analytics of large categorical data. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2849–2858, 2012.

[6] B. Alsallakh, E. Gröller, S. Miksch, and M. Suntinger. Contingency wheel: Visual analysis of large contingency tables. *Proc. EuroVA*, pages 53–56, 2011.

[7] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pages 52–60. IEEE, 1998.

[8] R. Arias-Hernandez, L. Kaastra, T. Green, and B. Fisher. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10, Jan 2011.

[9] F. Bendi, R. Kosara, and H. Hauser. Parallel sets: visual analysis of categorical data. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 133–140. IEEE, 2005.

[10] M. Bögl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch, and A. Rind. Visual analytics for model selection in time series analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2237–2246, Dec 2013.

[11] C. Borgelt. Efficient implementations of apriori and eclat. In *FIMI03: Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations*, 2003.

[12] J. D. Brender, M. U. Shinde, F. B. Zhan, X. Gong, and P. H. Langlois. Maternal residential proximity to chlorinated solvent emissions and birth defects in offspring: a case-control study. *Environmental Health*, 13(1):96, 2014.

[13] J. D. Brender, F. B. Zhan, L. Suarez, P. H. Langlois, and K. Moody. Maternal residential proximity to waste sites and industrial facilities and oral clefts in offspring. *Journal of occupational and environmental medicine*, 48(6):565–572, 2006.

[14] H. Brenner and M. Blettner. Controlling for continuous confounders in epidemiologic research. *Epidemiology*, 8(4):429–434, 1997.

[15] R. L. Brent. Environmental causes of human congenital malformations: the pediatricians role in dealing with these complex clinical problems caused by a multiplicity of environmental and genetic factors. *Pediatrics*, 113(Supplement 3):957–968, 2004.

[16] B. Broeksema, A. C. Telea, and T. Baudel. Visual analysis of multi-dimensional categorical data sets. In *Computer Graphics Forum*, volume 32, pages 158–169. Wiley Online Library, 2013.

[17] Z. Bursac, C. H. Gauss, D. K. Williams, and D. W. Hosmer. Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 3(1):17, 2008.

[18] S. L. Cessie and J. C. V. Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):191–201, 1992.

[19] R. Chaves, J. Ramírez, J. Górriz, C. G. Puntonet, A. D. N. Initiative, et al. Association rule-based feature selection method for alzheimers disease diagnosis. *Expert Systems with Applications*, 39(14):11766–11774, 2012.

[20] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

[21] L. Cohen, L. Manion, and K. Morrison. *Research methods in education*. Routledge, 2013.

[22] H. Crammer. Mathematical methods of statistics, princeton, 1946.

[23] R. Dawson. The unusual episode data revisited. *Journal of Statistics Education*, 3(3):1–7, 1995.

[24] J.-B. du Prel, G. Hommel, B. Rohrig, and M. Blettner. Confidence interval or p-value? *Deutsches Arzteblatt International*, 106(19):335–339, may 2009.

[25] W. Duch, R. Adamczak, and K. Grabczewski. Extraction of crisp logical rules using constructive constrained backpropagation networks. In *Neural Networks, 1997., International Conference on*, volume 4, pages 2384–2389. IEEE, 1997.

[26] M. Efroymson. *Multiple regression analysis*. Mathematical Methods for Digital Computers, Wiley, New York, 1960.

[27] G. Ellis and A. Dix. Enabling automatic clutter reduction in parallel coordinate plots. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):717–724, 2006.

[28] G. Ellis and A. Dix. A taxonomy of clutter reduction for information visualisation. *IEEE transactions on visualization and computer graphics*, 13(6):1216–1223, 2007.

[29] J. J. Faraway. *Linear models with R*. CRC Press, 2014.

[30] S. J. Fernstad and J. Johansson. A task based performance evaluation of visualization approaches for categorical data analysis. In *Information Visualisation (IV), 2011 15th International Conference on*, pages 80–89. IEEE, 2011.

[31] S. J. Fernstad, J. Shaw, and J. Johansson. Quality-based guidance for exploratory dimensionality reduction. *Information Visualization*, 12(1):44–64, 2012.

[32] I. K. Fodor. A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*, 9:1–18, 2002.

[33] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[34] M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994.

[35] M. Friendly. *Visualizing categorical data*. SAS Institute Cary, NC, 2000.

[36] E. Gilbert-Barness. Teratogenic causes of malformations. *Annals of Clinical & Laboratory Science*, 40(2):99–114, 2010.

[37] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[38] D. V. Glidden, S. C. Shiboski, and C. E. McCulloch. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer, 2011.

[39] L. A. Goodman and W. H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764, 1954.

[40] S. Greenland. When should epidemiologic regressions use random coefficients? *Biometrics*, 56(3):915–921, 2000.

[41] F. Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[42] F. E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, 2001.

[43] J. A. Hartigan. Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4(3):187–213, 1975.

[44] M. A. Hernán, S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology*, 155(2):176–184, 2002.

[45] H. O. Hirschfeld. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge Univ Press, 1935.

[46] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[47] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 3–10. IEEE, 2010.

[48] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[49] J. Jaccard. *Interaction effects in logistic regression*, volume 135. Sage, 2001.

[50] S. Johansson. Visual exploration of categorical and mixed data sets. In *Proceedings of the acm sigkdd workshop on visual analytics and knowledge discovery: Integrating automated analysis with interactive exploration*, pages 21–29. ACM, 2009.

[51] S. Johansson, M. Jern, and J. Johansson. Interactive quantification of categorical variables in mixed data sets. In *Information Visualisation, 2008. IV'08. 12th International Conference*, pages 3–10. IEEE, 2008.

[52] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):993–1000, 2009.

[53] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. Prenticee Hall, London, 2002.

[54] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[55] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[56] I. T. Jolliffe. *Principal component analysis*, volume 487. Springer-Verlag New York, 1986.

[57] D. A. Keim. Designing Pixel-Oriented Visualization Techniques: Theory and Applications. *IEEE Transactions on Visualization and Computer Graphics*, 6:59–78, 2000.

[58] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, 2010.

[59] S.-J. Ko and J.-H. Lee. Feature selection using association word mining for classification. In *International Conference on Database and Expert Systems Applications*, pages 211–220. Springer, 2001.

[60] E. Kolatch and B. Weinstein. Cattrees: Dynamic visualization of categorical data using treemaps. *Project report*, 2001.

[61] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE transactions on visualization and computer graphics*, 12(4):558–568, 2006.

[62] J. Krause, A. Perer, and E. Bertini. Infuse: interactive feature selection for predictive modeling of high dimensional data. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1614–1623, 2014.

[63] J. B. Kruskal and M. Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

[64] H. O. Lancaster and E. Seneta. *Chi-Square Distribution*. Wiley Online Library, 1969.

[65] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of massive datasets*. Cambridge University Press, 2014.

[66] M. Levandowsky and D. Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.

[67] F. Levine and M. Muenke. VACTERL association with high prenatal lead exposure: similarities to animal models of lead teratogenicity. *Pediatrics*, 87(3):390–392, 1991.

[68] M. Lichman. UCI machine learning repository, 2013.

[69] B. Liu, K. Zhao, J. Benkler, and W. Xiao. Rule interestingness analysis using olap operations. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 297–306. ACM, 2006.

[70] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[71] B. L. W. H. Y. Ma. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, 1998.

[72] C. Mann. Observational research methods. research design ii: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, 20(1):54–60, 2003.

[73] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *Visual Analytics Science and Technology, 2011 IEEE Conference on*, pages 111–120. IEEE, 2011.

[74] T. M. Mitchell. *Machine learning*. The McGraw-Hill Companies, 1997.

[75] T. Mühlbacher and H. Piringer. A partition-based framework for building and validating regression models. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):1962–1971, 2013.

[76] A. Oyefeso, C. Clancy, and R. Farmer. Prevalence and associated factors in burnout and psychological morbidity among substance misuse professionals. *BMC Health Services Research*, 8(39):1–9, Feb. 2008.

[77] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim. Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data. In *Computer Graphics Forum*, volume 31, pages 1345–1354. Wiley Online Library, 2012.

[78] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 89–96. IEEE, 2004.

[79] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[80] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '94, pages 318–322, New York, NY, USA, 1994. ACM.

[81] SAS Institute Inc. *SAS/STAT Software, Version 9.1*. Cary, NC, 2003.

[82] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 65–72, 2004.

[83] G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[84] C. Steed, J. Swan, T. Jankun-Kelly, and P. Fitzpatrick. Guided analysis of hurricane trends using statistical processes integrated with interactive parallel coordinates. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 19–26, Oct 2009.

[85] L. M. Sullivan. *Essentials of Biostatistics in Public Health*. Jones & Bartlett Learning, second edition, 2011.

[86] R. F. Tate. The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, 42(1/2):205–216, 1955.

[87] J. J. Thomas. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005.

[88] W. M. THORBURN. The myth of occam's razor. *Mind*, XXVII(3):345–353, 1918.

[89] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[90] A.-K. Torgalsboen. Full recovery from schizophrenia: the prognostic role of premorbid adjustment, symptoms at first admission, precipitating events and gender. *Psychiatry research*, 88(2):143–152, 1999.

[91] P. van der Putten and M. van Someren. CoIL challenge 2000: The insurance company case. Technical report, Leiden Institute of Advanced Computer Science, 2000.

[92] E. J. Wegman and Q. Luo. High dimensional clustering using parallel coordinates and the grand tour. In *Classification and Knowledge Organization*, pages 93–101. Springer, 1997.

[93] X. Yan. *Linear regression analysis: theory and computing*. World Scientific, 2009.

[94] J. Yang, D. Hubball, M. O. Ward, E. A. Rundensteiner, and W. Ribarsky. Value and Relation Display: Interactive Visual Exploration of Large Data Sets with Hundreds of Dimensions. *Visualization and Computer Graphics, IEEE Transactions on*, 13:494–507, 2007.

[95] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 105–112. IEEE, 2003.

[96] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the symposium on Data visualisation 2003*, pages 19–28. Eurographics Association, 2003.

[97] T. J. Ypma. Historical development of the newton-raphson method. *SIAM Review*, 37(4):531–551, 1995.

[98] F. B. Zhan, D. J. Brender, H. P. Langlois, and J. Yang. Air pollution-exposure-health effect indicators: Mining massive geographically-referenced environmental health data to identify risk factors for birth defects. US Environmental Protection Agency, 2015. Final report (2011-2015, 325 pages).

[99] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual clustering in parallel coordinates. In *Computer Graphics Forum*, volume 27, pages 1047–1054. Wiley Online Library, 2008.

[100] B. Zou, J. G. Wilson, F. B. Zhan, and Y. Zeng. An emission-weighted proximity model for air pollution exposure assessment. *Science of The Total Environment*, 407(17):4939–4945, 2009.

[101] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.