ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/221654479

Visual exploration of categorical and mixed data sets

Conference Paper · June 2009

DOI: 10.1145/1562849.1562852 · Source: DBLP

CITATIONS
6

READS
871

1 author:

Sara Johansson Fernstad
Newcastle University

25 PUBLICATIONS **465** CITATIONS

SEE PROFILE

All content following this page was uploaded by Sara Johansson Fernstad on 02 May 2014.

The user has requested enhancement of the downloaded file.

Visual Exploration of Categorical and Mixed Data Sets

Sara Johansson
Visual Information Technology and Applications
Department of Science and Technology
Linköping University, Sweden
sara.johansson@itn.liu.se

ABSTRACT

For categorical data there does not exist any similarity measure which is as straight forward and general as the numerical distance between numerical items. Due to this it is often difficult to analyse data sets including categorical variables or a combination of categorical and numerical variables (mixed data sets). Quantification of categorical variables enables analysis using commonly used visual representations and analysis techniques for numerical data. This paper presents a tool for exploratory analysis of categorical and mixed data, which uses a quantification process introduced in [16]. The application enables analysis of mixed data sets by providing an environment for exploratory analysis using common visual representations in multiple coordinated views and algorithmic analysis that facilitates detection of potentially interesting patterns within combinations of categorical and numerical variables. The effectiveness of the quantification process and of the features of the application is demonstrated through a case scenario.

Categories and Subject Descriptors

I.3.6 [**Computer Graphics**]: Methodology and Techniques—*interaction techniques*; I.5 [**Pattern Recognition**]: Miscellaneous

1. INTRODUCTION

In many research and application areas data sets including categorical variables or a combination of categorical and numerical variables (mixed data sets) are nothing unusual. Although several similarity measures exist that can be used for categorical data, such as the Jaccard coefficient [26], overlap and Goodall similarity [4], these are usually not as straight forward and general as similarities within numerical variables. Due to this, categorical data is often more difficult to visualize and analyse. Moreover, many commonly used visualization techniques, such as parallel coordinates [14, 28] and scatter plots, have been developed for visualization of numerical data and are hence based on the numerical similarity or distance between data items. Several visualization techniques developed for categorical data exist, but their effectiveness is often highly dependent on the structure of the data and on the analysis task. Moreover, they are often unable to represent mixed data sets including a combination of categorical and numerical variables.

One approach to overcome the difficulties involved in visualization of categorical and mixed data sets is to quantify the categorical data, representing the categories with numerical values, which enables analysis using the more general methods developed for numerical data. To avoid misleading the analyst into drawing incorrect conclusions when employing this approach, it is crucial to find a quantification that preserves the relationships within the data set.

In [16] an interactive quantification process was presented, which utilises the efficiency of algorithmic data analysis as well as making use of the knowledge of domain experts. This process identifies numerical representations that preserve relationships within the data and enables modification based on the analysis task and user knowledge.

This paper presents MiDAVisT (Mixed Data Analysis Visualization Tool), which is an application that has evolved from the process and interactive environment presented in [16]. MiDAVisT is an interactive tool for analysis of categorical and mixed data sets that provides a combined algorithmic and user controlled quantification of categorical variables, enabling analysis using both algorithmic methods and visual representations developed for purely numerical data sets. MiDAVisT also provides an interactive environment for visual and exploratory analysis where commonly used visual representations for numerical data are provided and combined with common algorithmic analysis methods to facilitate detection of patterns and relationships between categorical and numerical variables. The effectiveness and usefulness of the application is demonstrated through a case scenario where relationships within and between categorical and numerical variables in a mixed data set are identified using the features provided in MiDAVisT. The main contributions of this paper can be summarised as:

- An interactive application for user controlled quantification and analysis of data sets including a combination of categorical and numerical variables, enabling analysis based on relationships within all variables of the data set.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
VAKD'09, June 28, 2009, Paris, France.
Copyright 2009 ACM 978-1-60558-670-0 ...$5.00.

- An exploratory environment including multiple coordinated views, where visual representations and algorithmic analysis methods developed for numerical data are provided for exploration and pattern detection in mixed data sets.

The paper is organised as follows. Section 2 presents related research. In section 3 the quantification process is described and in section 4 the visual exploration environment of MiDAVisT is presented. Section 5 contains a case scenario that demonstrates the features and effectiveness of MiDAVisT. This is followed by conclusions and future work in section 6.

2. RELATED WORK

Several visualization techniques exist that are designed specifically for visualization of categorical data. Some examples being fourfold displays [9], where the cell frequencies of two-by-two tables are represented by quarter circles, mosaic displays and mosaic matrices [7, 8, 9], which represent multi-way tables with tiles whose sizes are proportional to the cell frequencies. Parallel sets [18] is a visual representation with a layout similar to parallel coordinates [14, 28] where the categories of a categorical variable are represented with a set of boxes whose sizes are proportional to the category frequency. In parallel sets the numerical variables of mixed data sets are represented by separating numerical values into bins. Further one example where the layout of parallel coordinates is used for categorical data visualization is presented in [13], where parallel coordinates are extended to avoid data overlay, meaning data items being concealed by other data items. This is achieved by spreading the lines over additional axes and by sorting the lines according to what categories they belong to in the adjacent axes.

These techniques all attend to visualization of categorical and mixed data, and are hence related to the approach presented in this paper. However they all suggest single visual representations, whereas the application presented in this paper focuses on quantification of categorical data and analysis using common methods and visual representations for numerical data, hence providing a more general and diverse environment for visual analysis.

A range of similarity measures exist for measuring the similarity between individual categorical data items. The most simple one being the overlap similarity [4] which assigns a similarity value of 1 if two items match for a variable and 0 if they do not match. Although straight forward, the overlap similarity has a major drawback in that all matches and all mismatches are treated as equal. The Jaccard coefficient [26] is a similarity measure for binary data which is also based on category matching, but only considers matching of ones for binary items, whereas matching of zeros is ignored, this makes the Jaccard coefficient a suitable similarity measure for sparse data. None of these similarity measures are, however, suitable for the application presented in this paper unless modifications are made, since categorical variables are, in general, not binary and since both measures only consider whether items match or not and do not take any other properties of similarity into consideration.

Another approach to similarity in categorical data is to use data-driven similarity measures, such as the Goodall, Occurrence Frequency, and Smirnov similarity measures [4]. For these measures the frequency distribution of variables is taken into account when measuring similarity and, as a result, the behaviour of the measures is directly dependent on the structures in the data set, and is hence not as general as numerical similarity.

Several approaches to quantification of categorical data have been previously presented. In [19] a technique for ordering of categorical data is introduced, where clusters of categories are formed based on domain semantics and the categories are ordered in a way that minimises the distances within the clusters. In [24] categorical data is quantified based on the association of categories in a categorical space. The quantification is achieved using Correspondence Analysis (CA) [12], as described in detail in section 3.1. In [21] this technique is incorporated into a framework for mapping of diverse data types.

CA has been used in different ways in visualization. In [7, 8, 9] it is used to reorder the categories in mosaic displays, and [12] presents a number of ways to visualize the result of CA using scatter plot techniques, such as CA Maps where CA is used to position the categories in a plot, and CA Bi-plots where each row and column of a table is displayed as a point. In [12] CA is also suggested as a technique for quantification of categorical data in order to apply statistical techniques that require numerical data.

The quantification approach of MiDAVisT is based on the quantification process presented in [16]. This process is similar to the approach presented in [24], but extends it by incorporating the relationships of numerical variables into the quantification process and by utilising the domain knowledge of expert users, as described in detail in section 3.

In addition to this MiDAVisT also provides an interactive environment for visual exploration by combining algorithmic analysis and multiple coordinated views. Using multiple coordinated views is a well established concept [2, 5, 23] that has been successfully used to overcome the difficulty of presenting large amounts of data in one screen while simultaneously making it possible to find detailed structures. A number of tool-kits and applications exist that can be used for visual exploration using multiple coordinated views combined with algorithm analysis, some examples are XmdvTool [27], GAV [15], InfoVis Toolkit [6] and the Hierarchical Clustering Explorer [25]. MiDAVisT has been implemented using the GAV framework.

3. INTERACTIVE QUANTIFICATION

This section briefly describes an interactive process for quantification based on algorithmic analysis and knowledge of domain experts, which was introduced in [16]. MiDAVisT employs this approach for quantification of categorical data, as well as for identification of similarities and relationships between categories. The data set used to demonstrate the process is an automobile data set containing 205 data items and including 6 categorical and 8 numerical variables [1].

When performing quantification of categorical data it is of high importance to find numerical representations that preserve the existing relationships within the data set. The quantification process described in this section uses CA to identify similarities between categories, however any algorithm able to identify similarities between categories could be used.

3.1 Correspondence Analysis

CA is a method for analysis of frequency tables, where each cell represent the frequency of a combination of categories [11, 12, 24]. An example frequency table for two categorical variables (eye colour and hair colour) is shown in table 1. For larger numbers of categorical variables, tables that contain all possible category combinations within the data set are used.

CA identifies similarities between the cells of the frequency table, and can be seen as a special case of Principal Components Analysis [17]. An example of similarity between categories can be seen in table 1 where the *brown* and *hazel* rows follow a similar pattern with highest frequency for *brown* hair and lowest frequency for *blond* hair. Hence *brown* and *hazel* can be considered as similar. The row representing *blue* eyes is less similar to *brown* and *hazel*, with high frequencies for both *brown* and *blond* hair.

Initially CA computes a correspondence matrix, $\mathbf{P} = \frac{\mathbf{N}}{\mathbf{n}}$ where $\mathbf{N}$ is the frequency table and n is the grand total of the table. $\mathbf{P}$ is normalised and centred (equation 1) using $\vec{r}$ and $\vec{c}$, which are vectors containing the row and column sums respectively, and $\mathbf{D_r}$ and $\mathbf{D_c}$, which are diagonal matrices with $\vec{r}$ and $\vec{c}$ as diagonals.

$$\mathbf{S} = \mathbf{D_r}^{-1/2}(\mathbf{P} - \vec{r}\vec{c}^T)\mathbf{D_c}^{-1/2} \tag{1}$$

CA identifies independent dimensions in the frequency table by applying Singular Value Decomposition (SVD), $\mathbf{S} = \mathbf{U\Sigma V^T}$ where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\mathbf{\Sigma}$ is a diagonal matrix where the diagonal values are singular values of $\mathbf{S}$ [10]. The first independent dimension explains most of the variance within the table, and the variance explained decreases with every succeeding dimension. Principal axes, $\mathbf{F}$, of the table rows are extracted as in equation 2.

$$\mathbf{F} = \mathbf{D_r}^{-1/2}\mathbf{U\Sigma} \tag{2}$$

Based on the theory of optimal scaling [12] the first principal axis of $\mathbf{F}$, which is related to the first independent dimension in SVD, can be used as numeric representations of the row categories in the frequency table. In this way a quantification is performed based on the relationships of all categories within the data set. Table 2 displays the first principal axis when CA has been applied to table 1. As can be seen the *brown* and *hazel* categories are represented by values close to each other, suggesting that they are similar, whereas *blue* is represented by a value further away, indicating that it is less similar to *brown* and *hazel*.

Table 1: A frequency table for two variables, eye colour (rows) and hair colour (columns). Each cell in the table represents the frequency of a combination of categories.

	Black	*Brown*	*Red*	*Blond*
Brown	11	20	4	1
Blue	3	14	3	16
Hazel	3	9	3	2
Green	1	5	2	3

Table 2: The first principal axis when CA has been performed on table 1. The values of this axis are used as numeric representations of the row categories.

	First principal axis
Brown	-0.5103
Blue	0.5516
Hazel	-0.2098
Green	0.1891

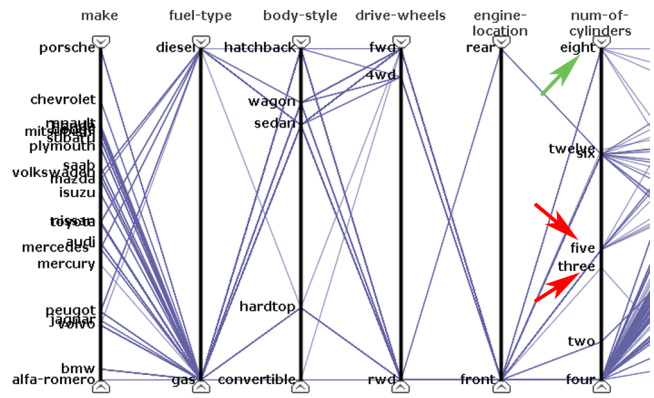

Figure 1: The suggested quantification of the categorical variables subsequent to correspondence analysis, displayed using parallel coordinates. The categorical variables are positioned to the left in the display and each category is represented by its name. The red arrows point out the categories *five* and *three* in the *number-of-cylinders* variable, which are considered similar to each other. A category that is considered less similar to them (*eight*) is pointed out by a green arrow.

3.2 Categorisation of Numerical Data

CA is performed on frequency tables where each cell represents the frequency of a combination of two categories, hence enabling a quantification based on the relationships within the categorical variables. If performing CA on a data set containing both categorical and numerical variables, the numerical variables need to be incorporated into the frequency table. This categorisation must be done in a way that preserves the existing numerical relationships. By this a quantification is performed that is based on relationships within both categorical and numerical variables.

The quantification process used in MiDAVisT provides two methods for categorisation of numerical data. One being a manual categorisation, where the user is allowed to interactively divide the data items into a number of categories guided by a visual display, and the other being an algorithmic categorisation using K-means clustering [20]. Using the K-means categorisation the relationships within the numerical variables are preserved, and hence the distance information within the numerical variables influences the quantification.

3.3 Interactive Modification

The quantification achieved through categorisation of numerical data followed by CA is presented to the user using

parallel coordinates, as shown in figure 1. This is a suggestion of how the categories can be quantified, based on the relationships within the data set, and also a presentation of similarities between categories. In figure 1 for instance, the categories *five* and *three* in the *number-of-cylinders* variable (pointed out with red arrows) are considered similar since they are positioned close together, whereas *eight* (pointed out with a green arrow) is considered less similar to *five* and *three* since it is positioned further away from them.

Although the algorithmic quantification is efficient, a domain expert may possess knowledge about the data and of the analysis task that the algorithm is unable to detect. To make use of this domain knowledge MiDAVisT provides possibilities for the user to modify the result of the quantification. The modifications include a manual reordering, which is performed by dragging and dropping categories within an interactive display, as well as a category weighting where the user assigns weight values to combinations of categories to indicate if they are to be more or less similar to each other, followed by CA re-computation as described in detail in [16].

In addition to this MiDAVisT also provides the possibility of undoing the modifications made by the user. This provides interactivity and exploratory freedom to the user by allowing analysis of different modifications without demanding a re-quantification to return to the quantification originally suggested by the algorithm.

3.4 Category Merging

For categorical variables with high cardinality, difficulties may arise if different categories are represented with numerical values close to each other. Since one category may conceal others this can cause a cluttered display. To avoid this MiDAVisT provide possibilities to merge several categories into one.

The merging can be performed based on the distance between the quantified categories, using a distance threshold which is interactively controlled by the user. This results in a merging where highly similar categories are merged into one representative category. Another approach available is a manual grouping where the user interactively selects a number of categories that are to be merged into one. In addition to this MiDAVisT provides possibilities of splitting any merged group of variables into its original categories throughout the analysis process. Figure 2 displays the scatter plot and figure 3 the graphical user interface (GUI) used for merging of categories. The glyphs of the scatter plot represent the categories of a selected variable, the x-axis represents the first principal axis achieved through CA, and the y-axis represents the second principal axis. In the GUI a threshold is set to identify categories that are close enough to be merged into one category, the suggested groups are represented by colour in the scatter plot, where the same colour is used for all categories within a group. One group containing three categories is selected and highlighted in black in figure 2, and the names of the categories are displayed within the GUI (figure 3).

4. VISUAL EXPLORATION

The quantification process described in section 3 not only provides a quantification of categorical variables which is based on relationships within the whole data set and on the

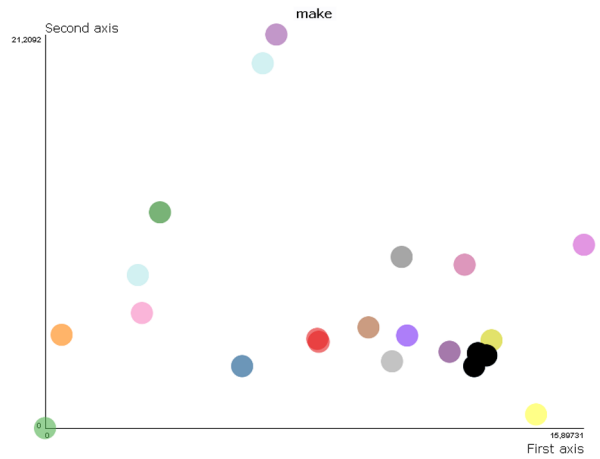

Figure 2: The scatter plot of the category merging interface, which displays the distribution of categories within a selected variable along the first and second principal axis resulting from correspondence analysis. In this example three categories (highlighted in black) are selected to be merged into one.

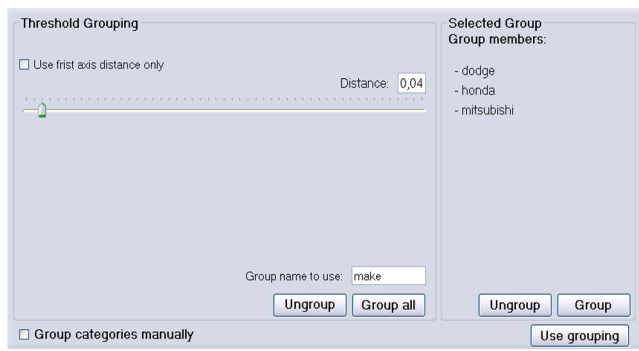

Figure 3: The graphical user interface (GUI) for merging categories. Merging can be either fully manual or guided by a similarity threshold. The names of the three categories selected in figure 2 are displayed in the right part of the GUI.

knowledge of an expert user. Through the visual representations used, it also provides an understanding of relationships and similarities between categories. Furthermore, any data set that has been quantified in MiDAVisT can be saved and re-opened at a later time, preserving all information on quantification and modifications made by the user.

In addition to the quantification MiDAVisT also provides an interactive environment for visual exploration of categorical and mixed data sets using multiple coordinated views. Within this environment the data set is treated as a numerical data set and three common visual representations for numerical multivariate data are used. Figure 4 shows the environment where a scatter plot matrix [3] is positioned in the top left view, a table lens [22] in the top right view and parallel coordinates [14, 28] in the bottom view. Within the table lens and parallel coordinates the category names are presented in addition to the numerical representations. The

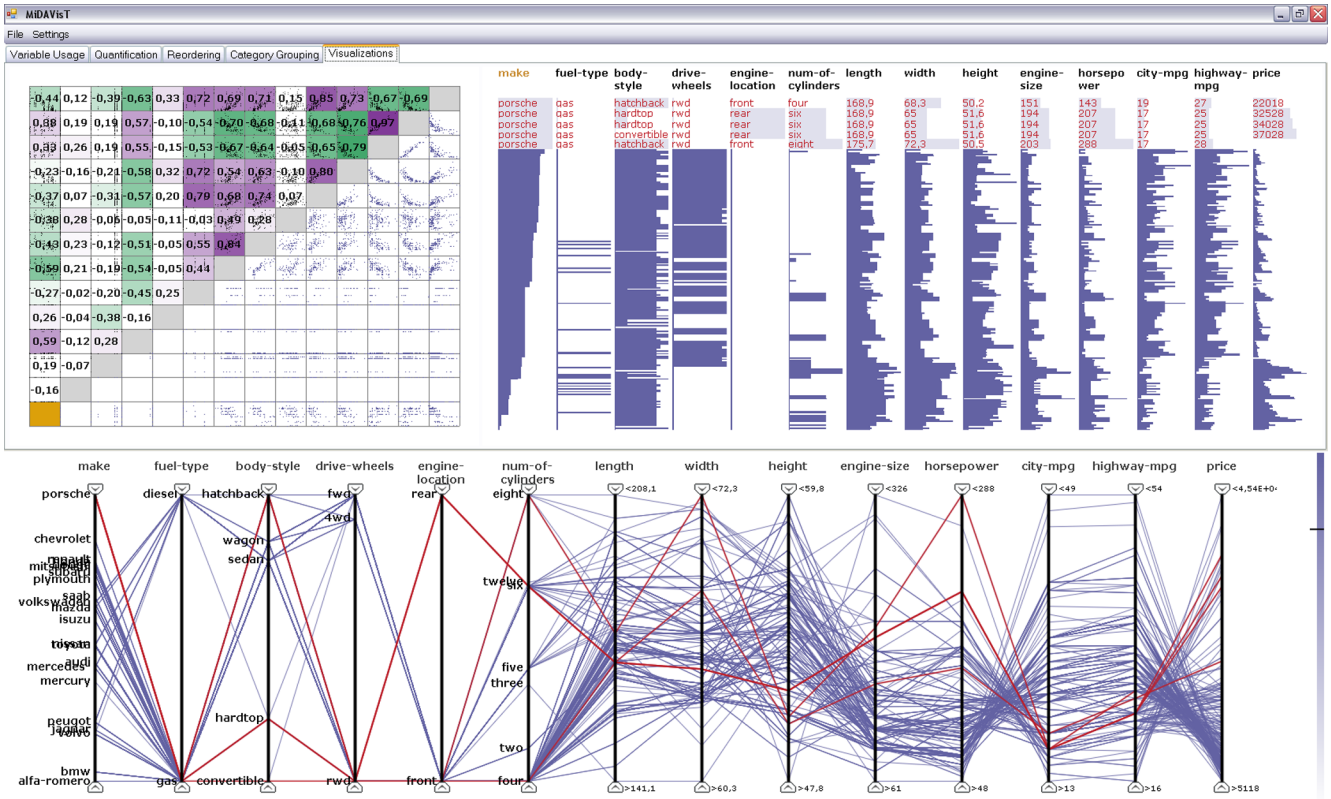

Figure 4: The multiple view environment for visual exploration of categorical or mixed data sets. In the top left view is a scatter matrix displaying variable pair correlation using green and purple colour. In the top right view is a table lens where the rows are ordered according to the *make* variable. The bottom view displays the data using parallel coordinates. The purple slider to the right of the parallel coordinates is used to control the transparency of the parallel coordinates lines. Colouring, selection and highlighting is coordinated between the views. In this example all Porsches are highlighted using red colour.

views are coordinated so that any selection or highlighting of items in one view is immediately reflected in the others.

To facilitate detection of structures and relationships in the data three different colour schemes are available within MiDAVisT. The default colour scheme uses a single colour to represent the data items that are not highlighted, as shown in figure 4. Using this colour scheme, individual items that are selected and highlighted in red are easily perceived in the parallel coordinates and table lens, enabling an understanding of the behaviour of individual items. The second colour scheme facilitates understanding of the relationships of categories, by supplying one individual colour for each category of a selected categorical variable, as shown in the top view of figure 5 where colouring is done according to the categories of the *number-of-cylinders* variable and parallel coordinates are used to display the colour schemes.

The third colour scheme, shown in the bottom view of figure 5, is used to emphasise cluster structures within the data set. This colouring is achieved by performing K-means clustering on the whole data set after the categorical variables have been quantified. Each cluster is assigned a unique colour and the data items are coloured according to their cluster membership.

In MiDAVisT the understanding of relationships between variables is facilitated through visual representation of correlation values. The Pearson correlation coefficient, r, is computed for every pair of variables according to equation 3, where N is the total number of data items and $\vec{x}_j$ and $\vec{x}_k$ are variables where $j, k = 1, ..., M$ and M is the total number of variables in the data set.

$$r(\vec{x}_j, \vec{x}_k) = \frac{N \sum_{i=1}^{N} x_{i,j} x_{i,k} - \sum_{i=1}^{N} x_{i,j} \sum_{i=1}^{N} x_{i,k}}{(N \sum_{i=1}^{N} x_{i,j}^2 - (\sum_{i=1}^{N} x_{i,j})^2)(N \sum_{i=1}^{N} x_{i,k}^2 - (\sum_{i=1}^{N} x_{i,k})^2)} \quad (3)$$

The correlation values are represented by coloured cells in the top left half of the scatter matrix, as shown in figure 4, where each cell represent the correlation of a variable pair. Positive correlation is represented by purple and negative correlation by green. Strong correlations are represented by darker colour than weak correlation.

5. CASE SCENARIO

This section describes how a fictional person, a veterinarian named Cate, can use MiDAVisT to analyse a mixed data set. The data set used in the case scenario is a slightly re-

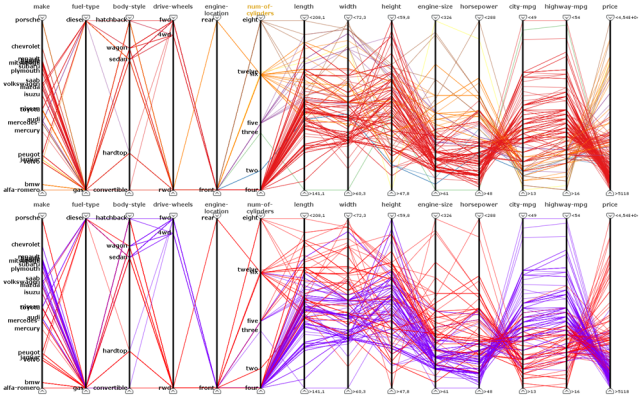

Figure 5: Two of the colour schemes available in MiDAVisT. The top view displays a colouring according to the *number-of-cylinders* variable, where each category of this variable is assigned a unique colour. The bottom view displays a cluster colouring, where colours are assigned according to cluster belonging when K-means clustering has been performed on the whole data set.

duced version of the horse colic data set in [1], including 13 categorical variables and 5 numerical. All conclusions drawn in the scenario that require domain knowledge are based on additional information included with the data set.

Cate is about to analyse a data set of 300 horses that have been treated for colic. The data contains information on different symptoms such as body temperature, pulse, abdomen shape and pain, as well as additional information on whether surgery was performed and on the outcome of the treatment. Cate has previously been introduced to some interactive tools for data visualization and exploration through a friend, and is hence familiar with common visual representations.

Cate starts the analysis by loading the data set into MiDAVisT, and decides to use the clustering approach to categorise the numerical variables, since she is mainly interested in relationships between symptoms that are based on all variables in the data set. MiDAVisT automatically performs K-means clustering followed by correspondence analysis, as described in section 3, and within a few seconds a quantification suggestion is presented using parallel coordinates (figure 6). Within this display Cate can for instance see that the *distended small intestines* and *distended large intestines* categories (pointed out with red arrows) are positioned close together, indicating that most of these horses had similar symptoms, whereas the horses that have a *normal* abdomen (pointed out with a green arrow) mostly have less similar symptoms to the horses with *distended small intestines* and *distended large intestines*, as indicated by it being positioned further away. Furthermore she can see that the quantification indicates that horses having *dark cyanotic* or *bright cyanotic* coloured *mucous membranes* (pointed out with bright blue arrows) have similar symptoms, which agrees with her knowledge that both these colours are indicators of serious circulatory compromises.

In general Cate considers that the suggested quantification and similarities agrees with her previous knowledge and experience. However, she finds the ordering of the *temperature of extremities* (third axis from left in figure 6) slightly illogical, since normal temperature is positioned closer to cold than to cool. Due to this she decides to manually change the ordering of this variable by dragging the cool category and dropping it on top of the cold category. This immediately swaps the positions of the two categories, as shown in figure 7.

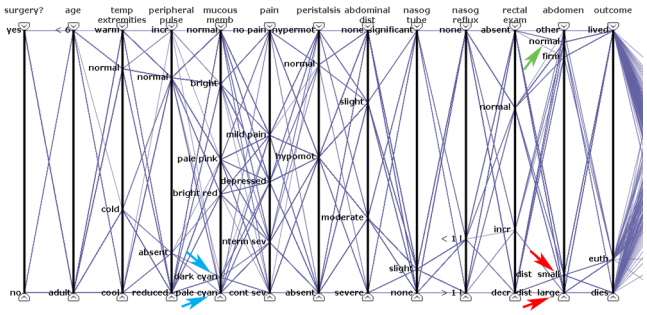

Figure 6: The suggested quantification of the horse colic data set. Categories positioned close to each other, such as *distended small intestines* and *distended large intestines* (red arrows), indicate that these horses have similar symptoms in general, whereas horses belonging to a category positioned further away, such as *normal* (green arrow), have symptoms that are less similar to the previous categories. The blue arrows point out the horses that have *dark cyanotic* or *bright cyanotic* coloured *mucous membranes*, which also are positioned close together, indicating similar symptoms.

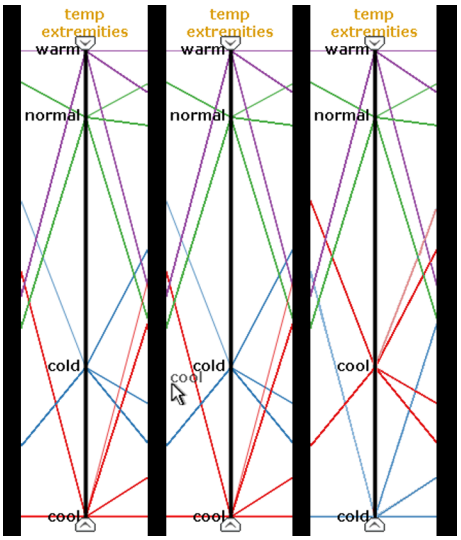

Figure 7: Manual modification of category positions. The left figure displays the suggested quantification. In the centre figure the user has dragged the *cool* category on top of the *cold* category. When dropping a category on top of another they swap positions, as shown in the right figure.

Figure 8: The multiple views environment when coloured according to the categories of the *outcome* variable, green represent *lived*, blue represent *euthanized* and red represent *dies*. In the table lens the rows are ordered according to the *abdominal distension* variable, with *none* as top category, followed by *slight* and *moderate*, and with *severe* as thin lines below. The empty lines at the bottom are missing values that represent horses where this variable was not recorded.

Since no variables contain large numbers of categories, Cate does not find it useful to merge any categories, and since she is now satisfied with the quantification she opens the visual exploration environment within MiDAVisT (figure 8). From her experience and knowledge Cate is aware that the abdominal distension is an important symptom, and based on this she decides to examine the relationship between the *abdominal distension* variable and the *outcome* variable. Figure 8 shows the multiple views environment when colouring is done according to the *outcome* variable, where green corresponds to horses that survived, blue to horses that were euthanized and red to horses that died from the colic. The rows of the table lens are ordered according to *abdominal distension* and, as can be seen, the two top categories in the table lens (which represent the *none* and *slight* categories) are mainly coloured green, whereas the lower categories, representing *moderate* and *severe* abdominal distension contain more red and blue. This indicates that there is a relationship between the *abdominal distension* of a horse and the *outcome* of the treatment.

One difficulty when displaying categorical variables using parallel coordinates is that the lines conceal each other, making it hard to get an understanding of the category frequencies. As a veterinarian Cate finds it important to know how many of the horses survived, which is hard to tell from the parallel coordinates. Due to this Cate reorders the rows in the table lens, using an ordering according to *outcome* instead of *abdominal distension*. Figure 9 displays the re-

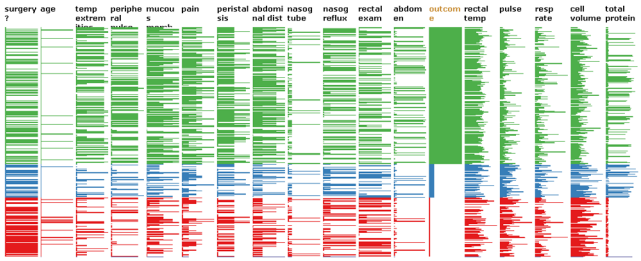

Figure 9: The table lens sorted according to *outcome*. With this ordering the category frequencies are easily perceived. More than half of the horses survived (green category), of the ones that did not survive less than half was euthanized (blue category), while the remaining died during the treatment (red category).

ordered table lens, where the frequency of a category can be told from the height of the group of rows that represent a category. In the figure the top category, representing horses that survived, is higher than the other categories, and hence most horses survived. From the other two categories it can be seen that, of the horses that did not survive, the majority were not euthanized (represented by blue).

In the scatter matrix (figure 10) Cate notices the group of purple cells in the bottom left part of the matrix. These

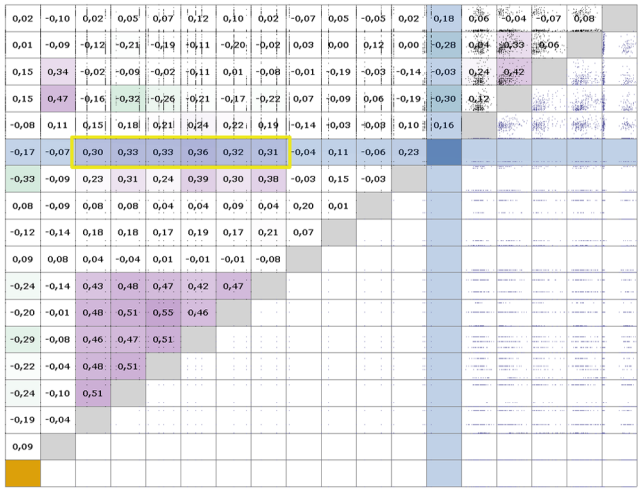

Figure 10: The scatter matrix of the quantified horse colic data set. In the bottom left part of the matrix a group of cells are coloured purple, due to high correlation between the variable pairs. The variable pairs including the *outcome* variable are highlighted in blue. The cells pointed out with a yellow rectangle are variables that have a correlation between 0.30 and 0.36 with the *outcome*.

cells represent a group of variables where all variable pairs are highly correlated. The variables are *temperature of extremities*, *peripheral pulse*, *mucous membrane colour*, *pain*, *peristalsis* and *abdominal distension*. The correlation indicates that all of these symptoms are related to each other. From the scatter matrix cells representing the *outcome* variable, which are highlighted in blue in the figure, it can be seen that the group of highly related variables are also correlated with the *outcome* variable, with correlation values ranging from 0.30 to 0.36, as pointed out by the yellow rectangle. This indicates that there is also a relationship between the previously mentioned symptoms and the outcome of the treatment, although not as strong as the relationship between the symptom variables.

The indicated relationships within the highly correlated group of variables agree with Cate's experience and domain knowledge, and she decides to continue her analysis to identify if there are any major groups of horses with similar symptoms and similar outcome. For this she decides to colour the data items using the clustering approach described in section 4. K-means clustering is applied to the whole quantified data set and the visual representations are coloured accordingly. The result, displayed using parallel coordinates, is shown in figure 11. From this it can be seen that two clusters exist, coloured red and purple, where the purple cluster mainly includes horses that survived whereas the red cluster mainly includes horses that were euthanized or died. By looking at the distribution of colours for the variables Cate is able to identify some interesting relationships that can be useful to her in her future practice and that verifies her previous experience. For instance she notices that most of the horses that did not survive had *pale pink*, *bright red*, *dark cyanotic* or *bright cyanotic* coloured *mucous membrane*, almost all of them had some *abdominal distension* and most of them also had high values for *packed cell volume*.

Cate is satisfied with what she has found so far, and decides to continue the analysis at some other time. She saves the quantification results for future analysis, and will be able to re-open the exploration environment using the same data set and quantification at a later time.

6. CONCLUSIONS AND FUTURE WORK

This paper presents MiDAVisT, an application for quantification of categorical data and exploration of data sets including both categorical and numerical variables. The quantification process used in MiDAVisT was introduced in [16] and enables a quantification based on the relationships within all variables of a mixed data set as well as utilising the knowledge of a domain expert.

MiDAVisT extends the analysis possibilities enabled by the quantification process by providing a multiple view environment for visual exploration and analysis of categorical and mixed data sets, using general and commonly used visual representations and analysis methods developed for numerical data sets. The main benefit of MiDAVisT is its ability to merge the quantification process into an interactive environment that enables versatile exploration.

The effectiveness of MiDAVisT is presented through a case scenario, where the quantification process as well as the features of the visual exploration environment are used to analyse symptoms and outcome of horses that were treated for colic. The scenario demonstrates how the quantification process can be used to identify symptoms that are closely related to each other, and how a mixed data set can be successfully explored by combining quantification and analysis methods traditionally used for purely numerical data, and how relationships between categorical and numerical variables can be identified through this.

Future work includes an evaluation of the efficiency and usefulness of the quantification process and the application together with domain experts and potential end-users. Furthermore additional analysis methods will be included in the exploration environment to further facilitate exploration and detection of patterns and relationships.

7. REFERENCES

[1] A. Asuncion and D. Newman. UCI machine learning repository. http://archive.ics.uci.edu/ml/, 2007.

[2] M. Q. W. Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the Workshop on Advanced Visual Interfaces*, pages 110–119, 2000.

[3] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, May 1987.

[4] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Siam International Conference on Data Mining*, pages 243–254. SIAM, April 2008.

[5] A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. In *Proc. IEEE Visualization ́91, San Diego, CA*, pages 156–153, 1991.

[6] J.-D. Fekete. The infovis toolkit. In *Proceedings of the 10th IEEE Symposium on Information Visualization*

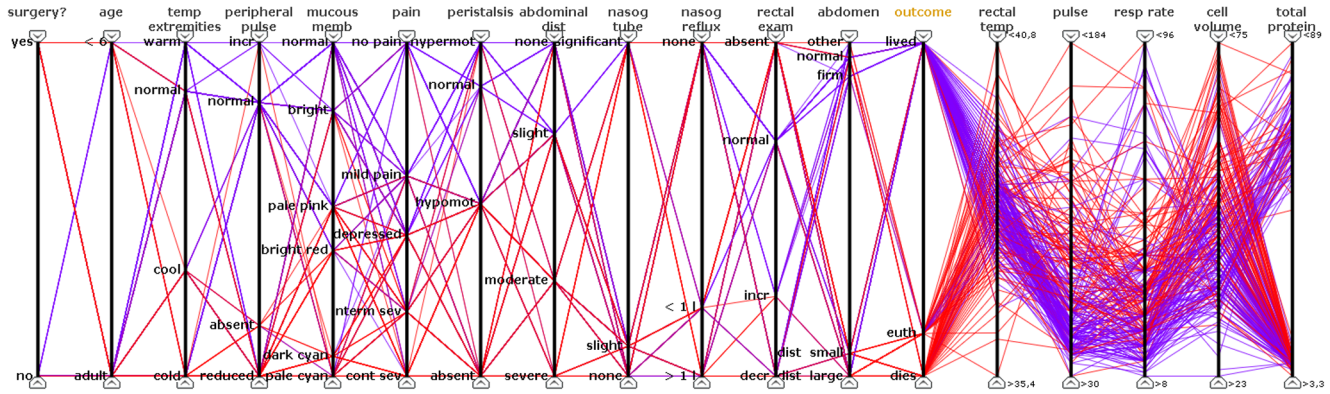

Figure 11: The parallel coordinates view in the exploration environment when the data is coloured according to a K-means clustering. As can be seen two clusters exist, coloured red and purple, where the red cluster mainly includes horses that died and the purple mainly includes horses that survived.

(InfoVis'04), pages 167–174. IEEE Press, October 2004.

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

[8] M. Friendly. Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3):373–395, 1999.

[9] M. Friendly. Visualizing categorical data: Data, stories, and pictures. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, April 2000.

[10] G. H. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *J. SIAM Numer. Anal.*, Ser. B(2):205–224, 1965.

[11] M. Greenacre. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall, 2006.

[12] M. Greenacre. *Correspondence Analysis in Practice, 2. ed.* Chapman & Hall, 2007.

[13] S. L. Havre, A. Shah, C. Posse, and B.-J. Webb-Robertson. Diverse information integration and visualization. In *Proceedings of SPIE - The International Society for Optical Engineering*, January 2006.

[14] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91, 1985.

[15] M. Jern, S. Johansson, J. Johansson, and J. Franzén. The gav toolkit for multiple linked views. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, CMV '07*, pages 85–97. IEEE Computer Society, July 2007.

[16] S. Johansson, M. Jern, and J. Johansson. Interactive quantification of categorical variables in mixed data sets. In *Proceedings of IEEE International Conference on Information Visualisation, IV08*, pages 3–10. IEEE Computer Society, July 2008.

[17] I. T. Jolliffe. *Principal Component Analysis, 2 ed.* Springer-Verlag, 2002.

[18] 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.

[19] S. Ma and J. L. Hellerstein. Ordering categorical data to improve visualization. In *IEEE Information Visualization Symposium Late Breaking Hot Topics*, pages 15–18, 1999.

[20] B. Mirkin. *Clustering for data mining a data recovery approach*. Chapman & Hall, 2005.

[21] A. Patro, M. O. Ward, and E. A. Rundensteiner. Seamless integration of diverse data types into exploratory visualization systems. Technical report, Worcester Polytechnic Institute, 2003.

[22] 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: celebrating interdependence*, pages 318–322. ACM, 1994.

[23] J. C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, CMV '07*, pages 61–71. IEEE Computer Society, July 2007.

[24] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang. Mapping nominal values to numbers for effective visualization. *Information Visualization*, 3(2):80–95, 2004.

[25] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results. *IEEE Computer*, 35(7):80–86, 2002.

[26] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*, chapter 2, page 74. Addison-Wesley, 2006.

[27] M. O. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proceedings of the Conference on Visualization 1994*, pages 326–333, October 1994.

[28] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of American Statistics Association*, 85(411):664–675, 1990.

View publication stats