A Memory-saving and Efficient Data Transformation Technique for Mixed Data Sets Visualization

Sun Yang, Zhao Xiang, Tang Daquan, Xiao Weidong
National University of Defense Technology, Changsha, Hunan, P. R. China
{victor_830514@yahoo.com, xiangzhao@nudt.edu.cn, tangdaquan@vip.sina.com, wilsonshaw@vip.sina.com}

Abstract
Although there have been effective visualizations for simplex continuous or categorical variables, mixed data sets are still difficult to visualize since no visualizations are available for them directly. This paper presents a memory-saving and efficient data transformation technique for mixed data sets visualization, particularly detailed on describing the application of correspondence analysis to quantify categorical variables, and proposed a set of cardinality reduction strategies to reduce the numbers of variables and values involved in computation. A series of empirical studies are carried out in a Star Coordinates-based environment to evaluate the visualization of mixed data sets based on transformed data sets. Finally we put forward the conclusion that the visualization results give a good graphical view of mixed data sets, and the proposed data transformation techniques are efficient in both time and memory.

Keywords--- Mixed data visualization, Data transformation, Clustering, Correspondence analysis, Cardinality reduction, Star Coordinates

1. Introduction
Existing in many application areas, the data sets that need to be explored often contain millions of objects described by tens, hundreds or even thousands of various types of attributes (variables). Different data types are used to define these attributes, including basic data types, e.g., integer, float, double, character, strings, and advanced data types, e.g., interval, ratio, binary, ordinal, categorical, etc. According to the semantics of attributes, one can always find a mapping between the related data types [1]. In this paper, we only concentrate on the mixed data sets that comprise two general data types, numeric and categorical, for other data types aforementioned can be mapped to one of these two types.

Mixed data sets are difficult to analyze since most existing visual exploration displays are designed to handle simplex numeric or categorical variables respectively, and that the difference between numeric or categorical data visualization tool is distinct. In order to reuse the existing visualizations, there are at least two main approaches to deal with mixed data sets. The first is to divide the continuous variables into bins, i.e. categorize the continuous variables, and display the data in categorical visualizations. However, the effect of visualization entirely depends on the domain knowledge of the analyst, and that blindly-transforming of numeric values into categorical displays may bring in artificial patterns, resulting in interpretation errors of the visualization. Moreover, visualizations designed for categorical variables usually only show the statistical information of the data sets well, but do not suffice for the display of distribution characteristic of data items. The second approach is to quantify the categorical attributes, i.e. to assign order and distance to the categories, as numeric values are known from categorical values by having a natural ordering and distance. The assigned order and distance among categorical values can convey notions as numeric variables do, such as magnitude and similarity. This approach takes advantage of the strongpoint of numerical visualization tools and helps the analyst have a quick exact discovery of distribution, connotative characteristics, relationships, patterns, trends and clusters of the target data sets. However, if the categories are quantified manually, the domain knowledge of the data analyst will be made use of inevitably, and simple transformation of the categorical values will bring the same problem as the first approach, producing artificial patterns. Thus, some algorithms have been proposed to compute distance and order of categorical data in mixed data sets. Johansson et al. [2] presented an interactive method to quantify the categorical value in mixed data sets, making use of the knowledge of the user. The quantification process employs clustering algorithm to categorize the continuous data first, and then uses Multiple Correspondence Analysis to compute the order and distance among categories, where information about relationships among the continuous variables is incorporated. At last, an interactive environment is
provided to the user, who is able to control and influence the computational process and analyze the results.

As far as we concern, however, there are at least two points need to be addressed in further research: (1) clustering is utilized to categorize continuous data, but no details about how to apply clustering is available, such as which continuous attributes to cluster; (2) for high cardinality variables categories are grouped together according to a user-set threshold value, applied to a scatter plot of the categories, which requires much domain knowledge. Still, there is not a complete solution to the problem that what if there are large amount of variables to process, which will result in great time and memory consumed. Inspired by the work of Johansson, this paper puts forward a refined approach to visualize mixed data, particularly detailed on applying clustering algorithms to categorize continuous variables, and proposed a set of strategies named cardinality reduction to reduce the number and cardinalities of variables involved in computation. In general, the main contribution of this paper can be concluded as:

- A full description of applying clustering to the categorization of continuous variables;
- A set of cardinality reduction strategies to preprocess the mixed data sets with a great many variables, as well as high cardinality variables; and
- A Star Coordinates-based visualization for efficiently visualizing mixed data sets.

The paper is organized as follows: Section 2 presents related research. Section 3 describes the data transformation technique used in the paper, including continuous variables categorization by clustering and categorical variables quantification using Correspondence Analysis. Section 4 proposes the cardinality reduction strategies and their application. Section 5 presents the empirical results, followed by conclusions and future work.

2. Related work

Several approaches were proposed to visualize categorical variables during the last decades. A common solution to display categories is to represent them by visual entities scaled according to their corresponding frequency. To show the relationships in a two-way contingency table, sieve diagrams [3] were designed, which seems to be good for presenting the information hidden in a data set with small number of categorical variables and values. While a large data set with many categorical variables and values would be difficult to handle in this manner. Similar to sieve diagrams, mosaic displays [4] adopts tiles that are proportional to the observed cell frequency to represent the counts in the contingency table, where a count is used to show how many cases surveyed contain certain values of two categorical variables. Theoretically, the mosaic matrix is able to accommodate large data sets with large numbers of categorical variables and values; practically, it shares the shortcomings of sieve diagrams, difficult to scale to visualize large data sets with multiple categorical variables and values in a single display. Fourfold displays [5] is specially designed to visualize data sets with only two categorical variables and with only two categorical values for each variable, which is the major limitation of this method. Adapted from Parallel Coordinates [6], Parallel sets [7] utilizes a set of boxes that are sized according to the category frequency to represent categories of a categorical variable. And mixed data sets is also supported by dividing the continuous variables into bins, but it works best with categorical data sets with only several variables which have few categories, otherwise split-up bars may become too small to visualized properly.

All the previous techniques are designed and developed to explore the relationships and associations among categorical data, which incorporate different visualization approaches to uncover relationships that may be hidden in the original data sets, and do not need to map the categorical values onto numeric values. Unfortunately, only the statistical and frequency information of the data sets are visualized without the distribution information. Besides, all those techniques do not handle data sets with large number of categorical variables, as well as high cardinality, which are yet commonly existed in application areas.

On the other hand, a number of excellent techniques designed to handle numeric variables have been proposed, such as Parallel Coordinates [6], Scatter Plots [8], Radviz [9], Star Coordinates [10], etc. They are capable of visualizing multivariate quantitative data well and can be used for exploring outliers, clusters, and relationship of the data. But the similar drawbacks they share are the limitations to deal with the categorical attributes such that they only work on numeric data well. When importing data sets with categorical values into such visualization tools, most solutions to date are rather simplistic, just treating categories as special cases with only a few values.

Possible solutions to map categorical variables into numbers include assigning order of categorical data by arbitrary ordering (e.g., alphabetical order) and ordering based on the value of another variable (e.g., time). However, these often create artificial patterns, thus equal spacing that does not convey the degree of similarity between categorical values is also assumed in ordering algorithms sometimes. Ma and Hellerstein [11] suggested an ordering technique for categorical data, and visualized them in scatter plots and parallel coordinates. They formed clusters of categories based on domain semantics and ordered the categories in a way that minimizes the distances within the clusters, in which only order was assigned to the categorical values, but no spacing. In case of the inherent intractability of the above mentioned approach, Beygelzimer et al. [12] presented an algorithm utilizing a spectral method. The previous two algorithms provide an elegant result for displaying datasets with a small number of categorical variables, but the scalability to large data sets with multiple categorical variables was not reported. Later, Rosario et al. [13]
quantified categorical data based on the distance and association of categories in a categorical space by incorporating Correspondence Analysis (CA) [14]. But CA is only able to quantify the categorical data, helpless on mixed data sets. To process the mixed data sets in a unified way, we firstly utilize clustering to categorize every numeric variable and then employ CA to the transformed data set to quantify the categories of each categorical variable. Through this way, the mixed data can be visualized in numerical visualizations finally.

3. Transformation of mixed data set

3.1. Notation

Definition 1: In a data set \( G(X) \), objects \( \{X_1, X_2, \ldots, X_n\} \) (\( n \) is the number of items in the set) are from the same domain, represented by the same set of attributes \( A_1, A_2, \ldots, A_m \). Each attribute \( A_i \) describes a domain of values, denoted by \( DOM(A_i) \). If attribute \( A_i \) is numeric, domain \( DOM(A_i) \) is represented by continuous values; If attribute \( A_i \) is categorical, the domain \( DOM(A_i) \) is defined as finite and unordered, e.g., for any \( a_{i1}, a_{i2} \in DOM(A_i) \), either \( a_{i1} = a_{i2} \) or \( a_{i1} \neq a_{i2} \), which contains only singletons.

Definition 2: An object \( X \) in \( G(X) \) is logically represented as a vector \( [x_{11}, x_{12}, \ldots, x_{1p}, \ldots, x_{n1}, x_{n2}, \ldots, x_{np}] \), where the first \( p \) elements are categorical and the rest are numeric, \( x_{ij} \in DOM(A_i), 1 \leq j \leq m \). And every object has exactly \( m \) attribute values, not allowing attributes to have missing values.

3.2. Correspondence analysis

Correspondence Analysis (CA) is a multivariate statistical technique, which is designed to analyze the association of categorical variables. And CA is similar to Principal Component Analysis (PCA) except that CA is for categorical variables while PCA is for numeric variables. There have been several versions of CA for different uses, such as Simple Correspondence Analysis (SCA), which is applied to two categorical variables, and Multiple Correspondence Analysis (MCA), which is employed to analyze more categorical variables. SCA takes a two-way contingency table like Table 1 as input, which contains the information about each categorical values and frequency of data items containing two certain values. Take \( a_{i1}, a_{i2} \) and \( c_{i1j} \) in Table 1 for an example, \( a_{i1}(a_{i2}) \) is one category (categorical value) in \( DOM(A_i)(DOM(A_j)) \), and \( c_{i1j} \) indicates the number of cases containing certain values of \( a_{i1} \) and \( a_{i2} \). While MCA is applied to a multi-way Burt table like Table 2, the meaning of which can be deducted according to two-way contingency table. Here we also assume that the variables within the first column of the table are target variables, and the variables within the first row of the table are analysis variables. Take Table 2 for an instance, variable \( A_j \) listed in the first column are named target variables; while, variables \( A_1, A_2, \ldots, A_j \), etc, listed in the first row are called analysis variables.

<table>
<thead>
<tr>
<th>Table 1 Two-way contingency table</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_{i1} )</td>
</tr>
<tr>
<td>( c_{i1j} )</td>
</tr>
<tr>
<td>( a_{i2} )</td>
</tr>
<tr>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

Table 2 Multi-way Burt table

| \( a_{i1} \) | \( a_{i2} \) | \( \ldots \) | \( a_{i3} \) | \( \ldots \) |
| \( c_{i11} \) | \( c_{i21} \) | \( \ldots \) | \( c_{i31} \) | \( \ldots \) |
| \( a_{i2} \) | \( c_{i12} \) | \( \ldots \) | \( c_{i22} \) | \( \ldots \) |
| \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) |

In this way, a multidimensional space can be defined by columns in a contingency or Burt table, and the rows can be regarded as points in the space. CA extracts new dimensions that are independent of each other in the space. Based on Optimal Scaling [13], each target categorical variable values can be mapped onto several scale values on the first independent dimension (the first principal axis) that explains most of the variance within the space, which is the output of CA.

3.3. Categorization of continuous variables

Since CA is only effective for categorical variables, the continuous variables within a mixed data set have to be categorized first before applying CA. At least two clustering possibilities can be suggested. The first is an interactive clustering assignment performed by the user. This clustering is achieved by assigning data items into clusters by each analyst. However, we can not require the user all to be professional, and that it is impractical if the data set is large, thus the second clustering possibility is to apply a self-acting algorithm to the continuous variables of the data set in order to turn aside domain knowledge. Clustering algorithm is adopted to be applied to each continuous variable \( A_j(p < i \leq m) \) so that each continuous variable is categorized into bins according to the distance among data items.

Known to all, k-means is quite efficient in clustering large numerical data sets, so we consider it as a good viable choice to divide data items into clusters by each continuous variable and categorize the variable into bins. However, any other numeric clustering algorithms can be
alternatives of k-means, such as BIRCH [15], CLARANS [16], though the performance may differ from case to case.

Through clustering, numerical data are grouped into clusters, whose values can be nominally represented by certain categorical values, named \(bin_{a_1}, bin_{a_2}, \) etc. Then the newly-computed nominal categorical values can be imported into the table together with the original categorical values in a unified style, forming tables like Table 3, i.e. the production of clustering is that the mixed data are transformed to be filled into Table 3, where contains the information about each categorical values and frequency of data items containing two certain values. The values here may refer to the nominal categorical values formed by clustering continuous variables. For an example, \(bin_{a_1}\) is a new category formed by applying clustering to continuous variable \(A_1\), \(c_{ik}\), indicates the number of cases containing certain values of \(a_i\), and belonging to the \(t^{th}\) cluster formed according to continuous variable \(A_2\) simultaneously; likewise, the meanings of other letters in the table can be deducted accordingly.

Table 3 Burt table after continuous variable categorization

<table>
<thead>
<tr>
<th>(a_{1})</th>
<th>(a_{2})</th>
<th>(\ldots)</th>
<th>bin_{k1}</th>
<th>bin_{k2}</th>
<th>(\ldots)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_{1})</td>
<td>(c_{11})</td>
<td>(c_{12})</td>
<td>(\ldots)</td>
<td>(c_{1k1})</td>
<td>(c_{1k2})</td>
</tr>
<tr>
<td>(a_{2})</td>
<td>(c_{21})</td>
<td>(c_{22})</td>
<td>(\ldots)</td>
<td>(c_{2k1})</td>
<td>(c_{2k2})</td>
</tr>
<tr>
<td>(a_{3})</td>
<td>(c_{31})</td>
<td>(c_{32})</td>
<td>(\ldots)</td>
<td>(c_{3k1})</td>
<td>(c_{3k2})</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
</tbody>
</table>

3.4. Quantification of categorical variables

After all continuous variables are categorized, MCA is applied to the Burt table like Table 3 and the quantified categorical values are computed from the first CA independent dimension coordinates using Optimal Scaling. And if the target variable is perfectly associated with analysis variable (e.g., one-to-many association or many-to-many association), modified optimal scaling proposed by Rosario [13] can be used as an alternative to Optimal Scaling. In this way, all categories in the data set will be transformed into numbers and numerical visualization tools, such as Parallel Coordinates and Star Coordinates, can be integrated to give a visual view of the transformed mixed data. The whole procedure for visualizing mixed data sets is over, providing users with a compact graphical display of the mixed data set.

4. Cardinality reduction strategies

Along with the increasing number and cardinality of variables in the mixed data sets, the memory-intensive and time-consuming problem of MCA will be more and more noteworthy. Rosario et al. proposed Focused Correspondence Analysis (FCA) [13] as an alternative to MCA to process a large number of categorical variables, some of which possibly have high cardinality. FCA does not simultaneously analyze all variables like MCA, which is less computationally efficient but more memory-saving when compared to MCA. And FCA analyzes the target categorical variable against its top \(k\) associated analysis variables instead of all the rest variables, resulting the lost of information that is originally contained in the unused variables. Moreover, FCA is not capable of correspondence analysis of mixed data sets directly. Aiming at computing the mixed data sets with a great many variables or high cardinality variables, here we present a set of strategies, which is referred as cardinality reduction, using less memory than MCA and being more efficient than FCA.

As mentioned in section 3.3, \(k\)-means is applied to each continuous variable, thus the number of analysis variables in Burt table equals the number of variable of the mixed data set. And the Burt table will have many columns if the mixed data set contains many continuous variables. In the cardinality reduction strategies, \(k\)-means directly works on all continuous variables, the output of which is a set of clusters according to all the continuous variables. Thus, the number of columns in the Burt table is reduced, forming a table like Table 4, which contains the information about each categorical values and frequency of data items containing two certain values. Here, the values may refer to the nominal categorical values formed by clustering all continuous variables. For an example, \(cluster_{a_1}\) is a new category formed by applying clustering to all continuous variable, and \(c_{i\{r\}}\) indicates the number of cases containing certain values of \(a_i\), and belonging to the cluster \(cluster_{a_1}\) simultaneously; like wise, the meanings of other letters in the table can be determined.

Table 4 Burt table of RCCA

<table>
<thead>
<tr>
<th>(a_{1})</th>
<th>(a_{2})</th>
<th>(\ldots)</th>
<th>(cluster_{1})</th>
<th>(cluster_{2})</th>
<th>(\ldots)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_{1})</td>
<td>(c_{11})</td>
<td>(c_{12})</td>
<td>(\ldots)</td>
<td>(c_{1r1})</td>
<td>(c_{1r2})</td>
</tr>
<tr>
<td>(a_{2})</td>
<td>(c_{21})</td>
<td>(c_{22})</td>
<td>(\ldots)</td>
<td>(c_{2r1})</td>
<td>(c_{2r2})</td>
</tr>
<tr>
<td>(a_{3})</td>
<td>(c_{31})</td>
<td>(c_{32})</td>
<td>(\ldots)</td>
<td>(c_{3r1})</td>
<td>(c_{3r2})</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
</tbody>
</table>
detailed accordingly. This kind of method is what we named \( r \)-Clustering based Correspondence Analysis (RCCA), where \( r \) drives at that it is a strategy implemented by clustering all the numeric variables.

If the mixed data mostly comprises categorical variables, we can use clustering algorithm for categorical variables to reduce the number of analysis variables of Burt table. The category items in the Burt table are replaced by the clusters formed according to all categorical variables. There are a lot of categorical clustering algorithms are available for use, and \( k \)-mode algorithm [1] is employed here for its efficiency and convenience. Then, the original long Burt table will be transformed to a table like Table 5, where \( \text{cluster}_1^r \), \( \text{cluster}_2^r \), \( \text{cluster}_3^r \), \ldots, etc, are formed by applying \( k \)-mode to all categorical variables. This kind of method is what we named \( c \)-Clustering based Correspondence Analysis (CCCA), where \( c \) drives at that it is a strategy implemented by clustering all the categorical variables. As combination of RCCA and CCCA, \( c\)-Clustering based Correspondence Analysis (CRCA) is proposed, which applies \( k \)-means and \( k \)-mode to continuous and categorical variables simultaneously and respectively, for RCCA and CCCA are independent from each other. And the Burt table of CRCA is Table 6, in which the meanings of letters can deducted accordingly.

Since simplex clustering algorithms are respectively applied to categorical variables and continuous variables to reduce the number of column in the Burt table, we can also group data items using mixed clustering algorithms against categorical and continuous variables simultaneously, such as \( k \)-prototype [1] and so on. This kind of method is what we named Mixed-Clustering based Correspondence Analysis (MCCA).

However, there are very few mixed clustering approaches, and all of them bear shortcomings as lack of accuracy, robustness and predictability. A recent finding in cluster ensemble [17] which has good stability and capability to deal with irregular or noise data, as well sound scalability could be a solution to combine multiple clusters without accessing the original features. Cluster ensemble is processed as follows: Clusterer \( \Phi_i \) is designed to each attribute \( A_i \) at first. If \( A_i \) is categorical, \( \Phi_i \) divides data items into partition \( \lambda_i \) based on the categorical value of \( A_i \), i.e.

\[
\lambda_i = \{ \Phi_i(X_i, x_i) | x_i \in \text{DOM}(A_i), X_i \in G(X) \} (1 \leq i \leq p);
\]

And if \( A_i \) is numerical, \( \Phi_i \) is \( k \)-means algorithm and partition data items into \( \lambda_i \). Then, the partitions of data set \( \lambda_1, \lambda_2, \ldots, \lambda_n \) are combine to \( \lambda \) using consensus function \( \Gamma \), which is defined as a function \( \mathbb{N}^{m^r} \rightarrow \mathbb{N}^p \) mapping a set of clusters to an integrated cluster: \( \Gamma: \{ \lambda_i | i \in \{1, \ldots, m\} \} \rightarrow \lambda \). If there is no a priori information about the relative importance of the individual groupings, then a reasonable goal for the consensus answer is to seek a clustering that shares the most information with the original clusters. Here maximal average mutual information is used as the objective function for cluster ensemble:

\[
\Phi_{\text{MAIM}}(\Lambda, \hat{\lambda}) = \frac{1}{r + p} \left[ \sum_{i=1}^{r} \delta \Phi_{\text{MID}}(\hat{\lambda}, \lambda_i) + \sum_{i=1}^{p} \Phi_{\text{MID}}(\hat{\lambda}, \lambda_i) \right],
\]

where \( \Lambda \) is the set of groupings \( \{ \lambda_i | i \in \{1, \ldots, m\} \} \), \( \delta \) is weighted value, \( \hat{\lambda} \) is one certain partition, \( \Phi_{\text{MID}}(\lambda_i, \lambda_j) \) is normalized mutual information:

\[
\Phi_{\text{MID}}(\lambda_i, \lambda_j) = \frac{2 \sum_{x \in \Lambda} n_{x\lambda_i} \log \left( \frac{n_{x\lambda_i} n}{n_x n} \right)}{n_x n},
\]

where \( k_x \) and \( k_y \) is respectively the number of clusters in \( \lambda_x \) and \( \lambda_y \), \( n_x \) is the number of objects in
cluster $c_a$ according to $\lambda_a$, $n_i$ is the number of objects in cluster $C_i$, $\lambda_b$, $n_{ab}$ denotes the number of objects that are in cluster $C_i$ according to $\lambda_a$ as well as in group $C_j$ according to $\lambda_b$. The optimal combined clustering is computed by optimizing the objective function:

$$\lambda_{opt} = \arg \max \left[ \delta \sum_{i=1}^{\delta} \Phi_{NMI}(\lambda_a, \lambda_b) + \sum_{i=1}^{\delta} \Phi_{NMI}(\lambda_a, \lambda_b) \right].$$

This kind of strategy is what we named Cluster Ensemble based Correspondence Analysis (CECA). The Burt table of MCCA and CECA is presented as Table 7, where $cluster_1$, $cluster_2$, $cluster_3$, etc. are formed by applying mixed clustering algorithms or cluster ensemble to the data set, and the frequencies indicate the numbers of cases containing certain categorical value, and belonging to certain corresponding cluster.

### Table 7 Burt table of MCCA and CECA

<table>
<thead>
<tr>
<th>cluster_1</th>
<th>cluster_2</th>
<th>cluster_3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$c_1(cluster_1)$</td>
<td>$c_1(cluster_2)$</td>
<td>$c_1(cluster_3)$</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$c_2(cluster_1)$</td>
<td>$c_2(cluster_2)$</td>
<td>$c_2(cluster_3)$</td>
</tr>
<tr>
<td>$a_3$</td>
<td>$c_3(cluster_1)$</td>
<td>$c_3(cluster_2)$</td>
<td>$c_3(cluster_3)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### 5. Empirical Results

In this section, the validity of the basic data transformation technique for mixed data sets visualization are assessed by comparing visual display of the MCA-based implementation and the arbitrary quantification (arbitrary ordering and uniform spacing)-based implementation. The memory-saving and time-efficiency of cardinality reduction are demonstrated by comparing MCA, FCA, RCCA, CCCA, CRCA, MCCA and CECA implementation. Here we extend the definition of MCA and FCA to include the processes of continuous variable categorization. In this way, FCA can be applied to Burt tables like Table 3 to process mixed data sets then.

The prototype system is developed with Java in eclipse 3.2 with Microsoft Windows XP SP3. Star Coordinates is chosen as the visualization tool, for it is an visual exploratory technique that is better in assisting users in discovering trends, outliers, and clusters in numeric data sets than Parallel Coordinates in the early stage of data-understanding tasks.

The data sets used in experiments are two real data sets, automobile and flag. Automobile data set includes 205 cases of various auto models, each of which has 15 continuous attributes, 10 categorical attributes and 1 integer attribute that can be mapped to categorical attribute. Flag data set contains 194 data items of national flags, each of which has 2 continuous attributes, 7 categorical attributes, 1 unique name, 8 integer attributes and 12 Boolean attributes. And the unique name, integer attributes and Boolean attributes can be mapped to categorical attributes. Data items that have missing values are omitted because our method does not involve cases like those. The two data sets are available at: [http://www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html).

### 5.1 Quality of visual display

Since the automobile data set is typical and easy to interpret, it is incorporated to compare the effects of our proposed basic data transformation technique and the arbitrary quantification approach for mixed data sets visualization. To make the comparison more obvious, we particularly analyzed 6 categorical attributes, i.e. make, body-style, drive-wheel, engine-type, num-of-cylinders and fuel-system, and 7 continuous attributes, i.e. wheel-base, highway-mpg, length, city-mpg, horsepower, price and engine-size.

![Figure 1 Visualization of arbitrary quantification](image-url)

Figure 1 is the initial phase of visualization without interaction using arbitrary quantification. The points representing the cases are distributed confusedly, and no connotative information of the data set can be obtained directly; while Figure 2 is the initial phase of visualization without interaction after applying our proposed data transformation. In the latter, the similarities and dissimilarities between categories are shown very obviously. Take the attribute make for an example, whose quantification result is presented in Figure 3, all the car makes are clearly partitioned into 5 groups. Mazda, Chevrolet, Dodge, Plymouth, Mitsubishi, Honda and Subaru follow a similar pattern, different from Mercedes-bens when taken the whole data set into consideration, so they are ordered close to each other but far away from Mercedes-bens. Moreover, a part of the
data items (in the red circle) cluster together and some connotative information can be revealed based on it. We can potentially forecast that it would be easy to discover more information accurately and quickly with further interaction and exploration.

Figure 2 Visualization of MCA-based quantification

Figure 3 Quantification of variable-make

5.2 Memory space and processing time

We agree with Rosario et al. [13] that the most memory-consuming part of the implementation is the application of CA. So the memory space needed of CA in every cardinality reduction strategies are focused and compared here. Because the MCA uses the most memory \((\text{sum of cardinality})^2\) by ignoring any specific memory optimization that may be implemented in CA, we illustrate the normalized memory space used by different cardinality reduction strategies in Figure 4, using

\[
\begin{align*}
\text{MCA} & \quad \text{MemorySpace} \\
\text{FCA} & \quad \text{RCCA} \\
\text{CCCA} & \quad \text{CRCA} \\
\text{MCCA} & \quad \text{CECA} \\
\end{align*}
\]

In the figure, it is clear that different cardinality reduction strategies have different effects on different kinds of data sets such that the effect of certain strategy may differ according to the constitutions of the data sets. For example, RCCA does not work well with the flag data set, even not better than FCA; while CCCA is quite memory-saving on the flag data set, because the flag data set has only two continuous variables, most of which are categorical. As for the automobile data set, in which the proportion of continuous variables and that of categorical variables are more or less the same, the capability of RCCA/CCCA is similar but better than FCA. And CRCA takes the advantages of both RCCA and CCCA, thus it is better than either of them. Further more, MCCA and CECA save the most memory than others during processing.

Figure 4 Memory-using of different strategies

Figure 5 Run times of different strategies

Figure 5 shows the normalized run time of different cardinality reduction strategies. In general, FCA does not analyze all categorical variables simultaneously and has more columns in the Burt table, both of which result in more computation time for CA. So we normalize run time of different cardinality reduction strategies using

\[
\begin{align*}
\text{MCA} & \quad \text{RunTime} \\
\text{FCA} & \quad \text{RCCA} \\
\text{CCCA} & \quad \text{CRCA} \\
\text{MCCA} & \quad \text{CECA} \\
\end{align*}
\]

Although cardinality reduction strategies, which cannot analyze all categorical variable simultaneously either, they are more efficient than FCA, because those cost less computation time in the process of CA due to the less columns in Burt table, and that the clustering algorithms are fast in most cases. However, RCCA is slower than FCA with the flag data set because cardinality of the set
just reduced a little after RCCA. As a combination of RCCA and CCCA, CRCA is time-efficient on both of the two data sets. MCCA is sometimes even more efficient than MCA, for its cost caused by that certain calculations cannot be reused in CA could be complemented by reducing columns of Burt table. In general, CECA have a bit greater run time than MCCA though they have the same Burt table as input, because the pre-process, i.e. cluster ensemble, will take more time in optimization than common mixed cluster algorithm do.

However, it is not recommended to adopt MCCA in cardinality reduction computation due to its limitation mentioned in Section 4, although MCCA uses less memory and time than other strategies. In a word, it is best for the user to choose one cardinality reduction strategy according to the constitutions of data sets or just use CECA/CRCA as a default choice.

Conclusions and future work

This paper presents a memory-saving and efficient data transformation technique for mixed data sets visualization. Firstly, a detailed description of the application process of CA to quantify categorical variables is presented. Then to deal with the mixed data set containing a number of variables, even some of which have high cardinality, a set of cardinality reduction strategies are proposed, including RCCA, CCCA, CRCA, MCCA, CECA. Finally a Star Coordinates-based visualization environment is provided for efficiently visualizing the transformed mixed data sets. We evaluate our approaches in terms of memory space requirement, run time, and quality of visual display. Our data transformation technique, which does not rely on domain knowledge, is better than the common practice of arbitrary quantification apparently. And in practical application, we suggest the user to choose a certain cardinality reduction strategy among RCCA, CCCA, CRCA, MCCA, CECA according to the constitutions of the data sets, because the performance of the strategies may differ from case to case. Besides, they can also be utilized for other fields, where fewer variables or lower cardinality of categorical variables are needed as input.

In the future, more specific evaluations will be conducted to prove the effectiveness and efficiency of cardinality reduction strategies. And more researches is planned to develop more user-friendly interaction and exploration, such as Focus + Context [18]. Moreover, how to select proper variables in the mixed data set for visualization to get more accurate connotative information is an open problem worth dedication.

Acknowledgements

We gratefully acknowledge our colleagues Zhou Cheng and Long Gucan for their assistance in the experiments during the research, as well as National Science Foundation of China for the funding of this research.

References


