Enhancing Interactive Visual Data Analysis by Statistical Functionality

Jürgen Platzer*

VRVis Research Center
Vienna, Austria

Abstract

Both information visualization and statistics analyse high dimensional data, but these sciences provide different ways to explore datasets. While techniques of the former field provide graphics so that the user can visually detect patterns of interest, statistical algorithms apply the capabilities of computers to produce numerical reports from the data. Based on this observation a combination of techniques of both sciences can help to overcome their drawbacks. For this purpose a library was compiled that contains statistical routines, which are of high importance for information visualization techniques and allow a fast modification of their results, to integrate possible adaptations in the interactive visual data mining process.

Keywords: Data Mining, Information Visualization, Clustering, Dimension Reduction, Outlier Detection

1 Introduction

The exploration of high dimensional datasets is a tremendously growing working field. With the capabilities of today's computers to handle data containing millions of data points and thousands of dimensions it is essential to apply efficient methods to extract the information the user is searching for. Statistical routines and techniques of information visualization are useful to achieve this goal. But as one method on its own has several shortcomings combinations between the different capabilities of these sciences could be developed to improve the exploration of multivariate data, the so called data mining process.

Information visualization techniques create graphics and animations that stress certain structures and aspects of high dimensional data. The user, who examines the data, applies his or her pattern recognition skills as well as the experience and knowledge about the data to draw the correct conclusions. This is an efficient approach to detect data items of special interest, examine the main trends in the data or investigate functional dependencies between variables.

In contrast to that statistical routines use the possibilities of computers, which execute millions of operations within milliseconds. This allows the fast calculation of facts and numerical summaries. Also models that can predict variable values or introduce a simplification of the data can be fitted. Thus coherences within attributes as well as significant patterns of the data items can be revealed and analysed.

Because of the usage of different systems that gather the information of interest, a combination of those sciences would introduce a verification of the results of the applied methods. Consequently an error detection approach for the data mining process could be established that decreases the probability that wrong conclusions are implied. But the intelligent application of the strengths of the disparate techniques also makes a more efficient data exploration possible.

To achieve this collaboration, in this work the implementation of a library containing statistical routines adapted for the use in information visualization applications is presented. Because of the vast number of routines developed in the field of statistics for data analysis and exploration, the basic functionality, that every visual data mining tool should provide, had to be determined. Furthermore aspects like robustness, which decreases the occurrence of distorted results caused by outliers, and fuzzyness, which allows soft decision boundaries to describe uncertainties, are considered. A further demand on the library is, that its routines must be able to process large datasets efficiently.

Additionally a sample application was developed to demonstrate possible combinations of visualization and statistics. In the focus of this tool are tasks like outlier detection, dimension reduction and clustering, where computational approaches are combined with visual verifications and user interactions that can manipulate the results of the statistical routine. Special attention was paid on an interactive workflow were the user can determine the order of the steps of the data mining procedure.

The following section outlines the main statistical approaches, that were considered for this work, and examples of applications that already realize an interactive collaboration between computational routines and visualizations. In section 3 the possible benefits of this combination approach are discussed, followed by a description of the implemented functionality for the statistics library. Afterwards a proof of concept case demonstrates the usefulness

*Jürgen.Platzer@VRVIS.at

of the integration of statistical methods into information visualization applications. Before the paper is concluded, issues considered for the implementation of the library are depicted in section 6.

2 Related Work

The information visualization techniques to illustrate multivariate datasets are manifold. They can be roughly classified into the four categories geometric projection techniques, icon-based and pixel-based approaches and finally hierarchical visualizations [13]. This work applies graphic representations of the first type, which maps the variables of the data on the screen space. The most popular approaches of this category are scatterplots and scatterplot matrices [8] as well as parallel coordinates [14]. For scatterplots two attributes are mapped on the x and on the y axis of the visualization space and data items are represented by points in the coordinate system that is spanned. To allow an illustration of all dimensions of a dataset a scatterplot matrix was introduced, which shows all possible tuples of variables by scatterplot visualizations. The parallel coordinates achieve the representation of all attributes by mapping them on axes which are drawn as equidistant vertical lines. The data items are illustrated by poly lines, which connect the projected dimension values.

Also the field of statistics provides a multitude of analysis procedures for data exploration. In the scope of this work only tasks are addressed that are of special importance for a visual data mining application. Therefore the multivariate outlier detection, dimension reduction and clustering techniques are considered.

As outliers strongly influence statistical routines and cause wrong results, an efficient detection of these object is crucial. For this purpose a variety of heuristics has been developed. The most popular approaches are distance based, density based and distribution based methods. Routines of the first type consider the distance of each data item to its k nearest neighbour. If this distance exceeds a user specified limit, the object is identified as outlier [17]. Density based techniques refer to a volume parameter and a minimum number of data points that has to be located within this volume to define dense regions. Data items that could not be assigned to such an area are marked as outlying [7]. The multivariate outlier detection application that is applied in this work uses a distribution based approach, which assumes that the data applies to a multivariate elliptic distribution. Therefore for each data item the robust distance is calculated, which is based on the robust estimate of the covariance matrix [21], that describes the shape of the data cloud. If the data objects correspond to the distribution constraint their distance measures show a chi-squared distribution. Consequently a chi-squared distribution quantile can be considered to determine a decision boundary that differentiates between outliers and actual data points.

Clustering approaches group similar data items to introduce partitions of the data. The main two methodologies for this task are hierarchical and partitional techniques. A hierarchical clustering based on a merging operations initiates each data item as cluster. Afterwards the two most similar clusters are merged to a new cluster. This procedure is iteratively performed until only one cluster representing the whole dataset remains. This nested group structure can be represented by the tree-like dendrogram. In contrast to that partitional approaches assign data items to clusters according to an update rule that optimizes a global energy function. The most popular algorithm of this type is the k means clustering [12], where k indicates the user defined number of partitions that are created. While these routines assign each data item to exactly one cluster, fuzzy clustering approaches exist, which calculate for each object membership values, that indicate to which degree it is associated with each cluster. For this purpose the fuzzy k means algorithm [6] was considered.

To reduce the dimensionality of a dataset three main techniques were introduced. The self-organizing maps (SOM) [18] are based on an unsupervised machine learning approach, that tries to iteratively fit reference vectors in data space to the structure of the data items. These vectors are connected in a two dimensional lattice which represents the low dimensional projection of the data. Multi dimensional scaling (MDS) techniques [19] try to achieve a projection of the multivariate data that maintains the distance relationships between pairs of data items. The simplest but nevertheless popular dimension reduction technique is the principal component analysis (PCA) [15], which evaluates the directions of the major variances in the data cloud. These directions are called the principal components, on which a mapping of the data items can be performed. As the first principal components describe the majority of the variance in the data, a subset of these artificial dimensions can be chosen to span a subspace containing the main information of the data space.

Feature subset selection approaches have the same aim as dimension reduction techniques. But a low dimensional representation of the data is achieved by choosing only the most informative data attributes. As this concept was developed for supervised machine learning routines, it is difficult to apply for data exploration, because no measure for the quality of an attribute can be intuitively introduced.

The integration of computational routines in information visualization applications gained importance in the last 10 years. Therefore mainly clustering and the creation of low dimension data representations were applied. The reasons why data partitioning has been favoured are that group finding algorithms provide a fast categorization of the data and significantly improve the detection and interpretation of the main trends. The focus on the reduction of variables simply rises from the fact that humans are used to think in three dimensional spaces, while multivariate datasets represent their main information in a higher number of attributes. To overcome this discrepancy projection methods as well as feature subset selection approaches were applied.

But while simple visualizations of statistical results only serve to explore and present them, an interactive combination of statistics and visual techniques is rarely realized. An example therefore is the Visual Hierarchical Dimension Reduction (VHDR) [24] system, which applies a hierarchical clustering on the attributes of the data. The introduced dimension groups can be investigated and modified by using InterRing [25] a radial visualization tool for hierarchical data. Finally representative dimensions per selected cluster can be chosen. This approach integrates the user's knowledge and experience into the feature subset selection task for which a starting point is created by a statistical routine.

An interactive visual feature subset selection and clustering tool is presented in [11]. By calculating a measure for the "goodness of clustering" for each pair of variables a color coded matrix visualization is established where bright fields represent attribute combinations that show significant cluster structures. As an ordering heuristic is applied on the attributes the user can identify light regions in the visualization which represent groups of variables that allow a partitioning of the data items. These dimensions can be selected and used for a hierarchical clustering approach, that detects groups of arbitrary shapes, because of the integration of graph and density based clustering concepts. The additional parameters, that were introduced by this enhancements, as well as the number of detected clusters in the data can be steered by interactive visualizations.

A further example describing the power of the combination of visualizations and computationl routines is the HD-Eye approach [13], which adapts the OptiGrid clustering [16], so that the user is involved in the group finding process. Therefore a density estimation of the clustering procedure is used to decide, whether the data space can be subdivided by seperators like hyper planes, that are positioned in sparse regions. To achieve this a set of projections is suggested from the system, for which an icon-based visualization indicates, if a mapping is helpful to decide, where a separator can be introduced. The user can select the projection that shows the most significant gaps between groups of data. A histogram-like view, showing the agglomerations of data items by high bars is applied to define a separator. This approach is iteratively applied, until no subspaces can be divided anymore. Consequently this approach is an attempt to incorporate the capabilities of the human visual system into a clustering routine to achieve better results.

3 Integration of Statistical Functionality in Visualization

As the sciences visualization and statistics rely on different systems that analyse the data, their weaknesses and strengths are mostly dissimilar. Interactive visual applications provide graphics that can be modified by the user to achieve an efficient information drill down process, where firstly an overview is given. Afterwards zooming and filtering techniques allow the concentration on patterns or data items of special interest. Finally details-on-demand operations show numerical summaries or the data values of the selected subset themselves. Therefore mainly the user's extraordinary pattern recognition skills and knowledge about the data guides the exploration process [23].

Contrary to that statistical routines are algorithms and calculations of formulas that use computers to cope with the enormous computational effort for large datasets. This implies that the applied procedures have to be well chosen for the data that should be analysed. Consequently if a dataset contains clusters of arbitrary shapes, a k means clustering may produce low-quality results, because it only creates spherical groups. As this example shows, a general purpose method may fail on a given dataset and the detection of this failure is difficult to accomplish. Furthermore the presence of outliers can also significantly distort results of statistical routines.

Therefore the following discussions propose possible combinations of statistical methods and information visualization techniques for clustering, outlier detection and dimension reduction that may compensate the drawbacks of individual approaches.

3.1 Grouping of Data Items

The most popular statistical routine in data mining applications is clustering, that introduces partitions of the dataset. The detected groups can be seen as a simplification of data that allows an easier interpretation of the main patterns. But clustering results are also used to create clearer visualizations and a reduction of data items for time consuming calculations. An example is shown in figure 1 were the dataset UVW [2] containing 149769 data items is illustrated in parallel coordinate view. The first visualization simply plots all data items, which results in a clutterd graphic, while the second incorporates the information of the fuzzy clustering approach and thus reveals the structure of the data.

Consequently a clustering algorithm can introduce a meaningful division of the data into groups. But a visual verification of these partitions is crucial, because the introduced clusters may not be appropriate for the structures in the data. As general purpose clustering algorithms suffer that the number of clusters has to be set and/or the created clusters show a specific shape their results could be manipulated to achieve a better fit of the real groups in the

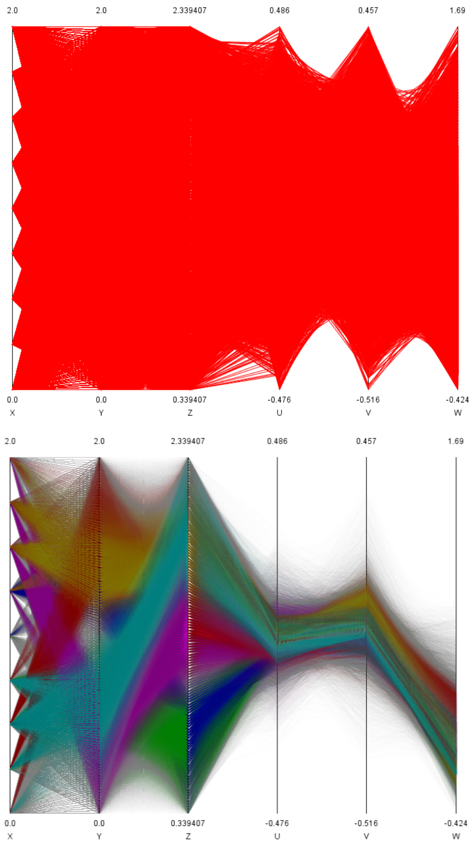

Figure 1: A cluttered parallel coordinate visualization (above) can reveal structure if clustering information is integrated (below).

data.

A visualization system that captures both the high dimensionality of the data as well as local features has to be applied to provide a user interface for the exploration and manipulation of clustering results. In the scope of this work the use of parallel coordinates and scatterplots is suggested. While the latter makes the intuitive investigation of two dimensional features possible, a parallel coordinates view illustrates all dimensions of a dataset. Furthermore dimension reduction techniques are applied to map the data items in a two dimensional space, which is visualized by a scatterplot. This allows a validation of the quality of the introduced partitions and represents a user interface for multivariate modifications of the clustering result.

As operations, that adapt the introduced partitions, clusters can be split, merged or deleted. Furthermore a cluster can be selected for a subclustering procedure, where only the data items of the chosen partition are considered for a clustering. But also cluster centers can be repositioned and objects can be reassigned to the cluster with the nearest center. After those interactions took place a reclustering based on the adapted clustering result can be initiated to improve the solution. Thus an interactive information exchange between a computational routine and the user's interaction is established, which is a significantly improved system in comparison to information visualization applications that only allow the exploration of clustering results. Because now the user is not restricted to the initiation of interactions, that are based on the perceived (mostly lower dimensional) features, also a routine that considers patterns in data space can be interactively applied.

3.2 Dimension Reduction and Feature Subset Selection

Based on a user defined similarity measure between attributes, a clustering procedure can be initiated to introduce groups of similar variables. Therefore a hierarchical clustering approach is adequate, because it allows the interactive modification of the group number. The established hierarchy of dimension relationships can be used as starting point for an interactive feature subset selection application that can also be combined with dimension reduction techniques. Thus a visualization of the dendrogram structure allows an interactive exploration of the clustering result. Dimensions that are represented by a selected node can be illustrated by parallel coordinates and serve as decision guidance for the feature selection. Additionally if for a group no representative dimension can be chosen, a dimension reduction approach is available. Consequently the main information represented by the cluster of dimensions is captured by a small number of artificial attributes, which can be selected.

The concept of this approach is shown in figure 2 by a simple dataset containing four attributes, from which two pairs are highly correlated. The parallel coordinates view on the left side illustrates this issue, while a representation of the cluster tree structure in the upper right corner shows that the clustering divides the attributes into two subgroups.

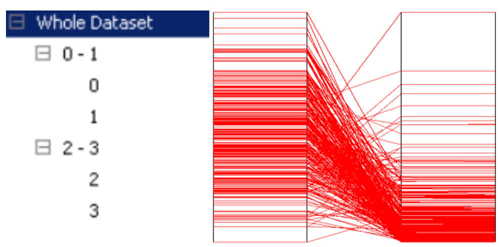

Figure 2: An attribute hierarchy representation and a parallel coordinate visualization showing the dimension relationships.

Because a clustering approach, that is not adapted to a specific kind of data, can produce arbitrarily bad fits to the structure of the dimension coherences, it is crucial that the user examines the achieved grouping. Therefore the most characteristic attributes of the clusters as well as those variables that can also be assigned to different partitions have to be determined. Dimensions of the first category may be those candidates that are chosen by the user to represent other dimensions for further operations. Other attributes may be of minor importance from the clustering point of view. Nevertheless the user can incorporate her/his knowledge in the process and can also choose those attributes, if they represent essential information.

Consequently this approach combines a statistical procedure to create an initial solution for the subset selection problem by introducing groups of dimensions for which a representative attribute can be chosen. But also the input of the user is required to choose the correct subset by visually validating the quality of the clustering. If the dimension clusters are visually heterogeneous, a new clustering can be tested to achieve a better result or a dimension reduction approach can be applied.

3.3 Multivariate Outlier Detection

In contrast to the detection of clusters, where similar data items are grouped, the identification of outliers searches for objects that deviate from the main behaviour of the data. Consequently this subset of data points may be heterogeneous. As browsing and selection techniques of information visualization applications only highlight data items showing similar properties, this approach is not adequate to detect high dimensional outliers. Therefore a statistical routine could be used again as an initial solution. In general such a method provides parameters that steer the number of identified outlying objects. Thus it is crucial to have a visual feedback that allows the interactive determination of the optimal parameter settings. To achieve this different linked visualizations, which are also able to apply dimension reduction techniques, could be used. A projection of the data items on a low dimensional subspace to realize a scatterplot illustration is especially helpful, because this approach allows the verification, whether the detected objects are at the border of the data cloud or deviate from the main groups in the dataset. Thus a validation of the statistical outlier detection is achieved and data items that are wrongly marked can also be manually deselected, which enhances the quality of the outlier detection.

The application of multivariate outlier exclusion is crucial for non-robust statistical routines, that calculate misleading results in the presence of outlying objects. In contrast to that visualizations of large datasets mainly stress the major patterns in the data. Thus it is essential to detect possibly outlying data items to accentuate their representations.

4 Library for Statistical Functionality for Visualization

For the determination of statistical functionality that is of high importance for information visualization applications software packages like SpotFire [5], Miner3D [4] or GGobi [1] were examined. Also publications of recent years, that discuss the integration of computational approaches in the visual data mining process were analysed. This research showed that the majority of the applied algorithms are concerned with clustering and dimension reduction. But also the use of transformations to prepare the data for further procedures was demonstrated especially in GGobi. Besides of these main tasks standard calculations like statistical moments and correlation measures were common.

In the scope of this work also a stronger integration of robust methods should be obtained. This is shown by implementing robust estimators for the location and the spread of a set of data values as well as by providing the calculation of robust correlation measures. But the main application which demonstrates the capabilities of robustness is the statistical outlier detection, which introduces a measure of outlyingness for each data item.

Furthermore the concept of fuzzyness is considered, because decisions made in the real world, from where the data comes from, are rarely reduced to yes/no answers. Therefore the fuzzy k means clustering was implemented, to show that it is not possible to assign each data item strictly to one cluster. Thus a degree of uncertainty is introduced to differentiate between objects that are near a cluster center and those data points that are located at cluster boundaries.

After introducing the main categories of functionality that is made available by the library, the remainder of this section gives an overview of the provided routines.

4.1 Transformations and Moments

Transformations can be seen as mappings of the data values to a certain interval or as modifications of the distribution of a set of values. The former application is useful to prepare a dataset for clustering, so that each dimension has the same range of values, which avoids that one attributes has a stronger influence on the distance calculations in the group finding process. The latter is of importance for statistical routines like the distribution-based high dimensional outlier detection, which can only be applied on data from a multivariate elliptical distribution. Therefore the statistics library provides linear, logarithmic, exponential and squareroot transformations.

As these mappings are applied on single dimensions separately also the statistical moments are in general calculated for attributes of the dataset. Therefore classic as well as robust estimates for the location (arithmetic mean, median, α - trimmed mean) and the spread (standard deviation, median of absolute deviations, inter quartile range) are provided.

4.2 Correlations and Covariances

To analyse the coherence between two variables three correlation measures are implemented. The classic Pearson correlation, which is biased if outliers are present, and the robust Spearman and Kendall correlations can be calculated. The two robust estimates do not only detect linear relationships but also exponential and logarithmic dependencies between dimensions.

A rough estimate for the shape of the multidimensional data cloud is given by the covariance matrix, which is a symmetric matrix holding the variances of the attributes in the main diagonal and the covariances between the dimensions in the off diagonal entries. As the covariance matrix describes the data as an hyper-ellipsoid it can be applied to integrate the shape of the data distribution into the distance calculation. This is achieved by the Mahalanobis distance. If a robust calculation scheme for the covariance matrix like the minimum covariance determinant (MCD) [21] estimator is applied, this concept can be used to calculate robust distances that assign high values to data items that strongly deviate from the majority of data items.

4.3 Clustering and Dimension Reduction

For the division of datasets into partitions the popular clustering procedures *k* means and fuzzy *k* means were implemented. While the first algorithm introduces a hard cluster structure, where each data item is assign to exactly one group, the fuzzy approach calculates memberships, that indicate to which degree an object belongs to a given cluster. Thereby the sum of the memberships for a data item always accounts 1. Additionally a hierarchical clustering approach based on the correlation matrix is realized to introduce groups of dimensions. This routine can be used as basis for a interactive feature subset selection application.

As dimension reduction routine the principal component analysis (PCA) is provided. It is based on the covariance matrix calculation. Thus also a robust PCA can be accomplished, by using the MCD estimate as covariance matrix.

4.4 Distributions and Statistical Tests

As theoretical distributions the normal, log normal, exponential, uniform and chi-squared distribution are realized. For each of these distributions values of the probability density function (pdf) and of the cumulative distribution function (cdf) as well as quantiles and random numbers are available. Additionally a Kolmogorov-Smirnov test can be performed to validate, whether a set of values comes from these theoretical distributions. Furthermore the invocation of tests, whether two attributes show the same distribution, is possible.

4.5 Regression

A least squares linear regression that predicts the values of a dependent variable based on a set of independent attributes is implemented. This model fitting approach is convenient for the identification of functional dependencies between dimensions of a dataset.

5 Proof of Concept Case

To demonstrate the advantages of the combination between computational routines and a visual verification, which also considers interactive modifications of the computed results and possible restarts of the algorithm, the interactive clustering tool that was realized for the sample application is presented. For this program the graphical user interface was based on the GTK+ [3] library, while the visualizations were implemented in OpenGL [22].

The interactive clustering applies a *k* means algorithm and illustrates its results in a scatterplot visualization of the data items mapped on the first two principal components. The cluster centers are also illustrated and used to manipulate the introduced partitions. Recluster processes are based on the repositioned cluster centers. Additionally scatterplots and parallel coordinates make the inspection of the original data possible.

As dataset the letter image recognition data [10] containing 20 000 observations and 16 numeric variables, which describe the properties of letters given as black-and-white images is used. For this proof of concept case only the letters A, B, C, D, E and F are considered, which reduces the number of data items to 4640. For the clustering procedure all numeric attributes were used except the horizontal and vertical position of the bounding box of the letters that do not contain information to discriminate the letters. As the dataset is based on the discretization of the letters by a pixel raster, the measurements show integer values and are not continuous.

Before the clustering was initiated the value ranges of all dimensions were mapped to the unit interval. As the first and the second principal component capture 48 % of the variance in the data, the visualization of dimension reduction represents a good hint for the structures in the data. Certainly an enormous amount of information is not considered, and thus not illustrated in the scatterplot visualizations shown in the figures 3, 5 and 7.

The interactive clustering process starts with an initial clustering, which introduces partitions that can be explored and manipulated. The result of this initial step is shown in the figures 3 and 4. While the former illustration in the scatterplot provides an intuitive overview of the data and its multidimensional structures, the parallel coordinates view makes the investigation of the partition shapes on single attributes possible.

The overview of the data, which is realized by the dimension reduction, shows that the green cluster, which

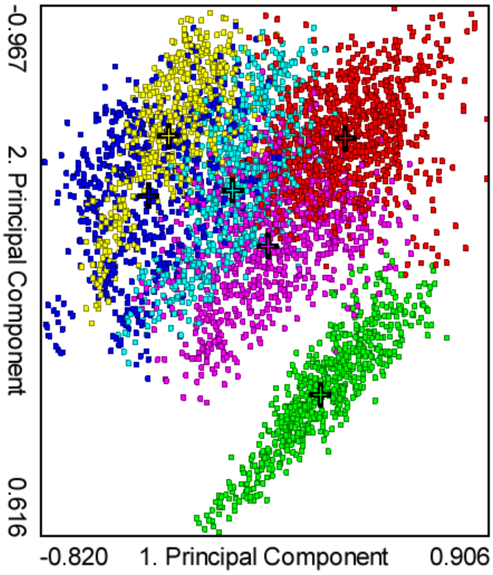

Figure 3: Clustering result mapped on principal components

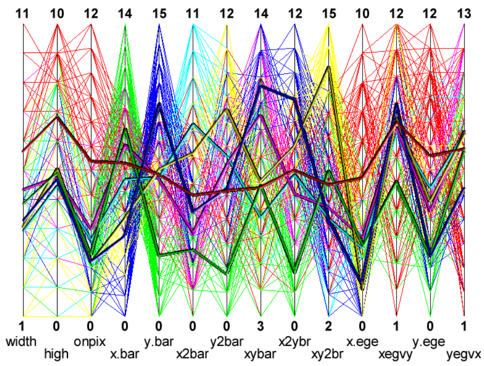

Figure 4: Clustering result in parallel coordinates

represents measurements of the letter A is very good separated. The remainder of the data can not be partitioned that easy. In the parallel coordinates it can be observed that the cluster centers show deviations in specific dimensions from the main behaviour of the data. As example the red cluster shows significantly higher values in the first three depicted dimensions, while the center of the green partition is discriminated by low values in the attributes *y.bar*, *y2bar* and *x2ybr*.

To emphasize on the differences between the clusters their centers have been manually repositioned in the scatterplot visualization shown in figure 5 so that the clusters focus on certain areas in the data. The modifications are also projected back into the data space and visualized in the parallel coordinate view (figure 4). There it is obvious that the modifications assigned more one dimensional extreme values to the cluster centers. To verify if this manipulation results in a better discrimination of the clusters a reclustering, taking the repositioned centers as initial solution, is initiated.

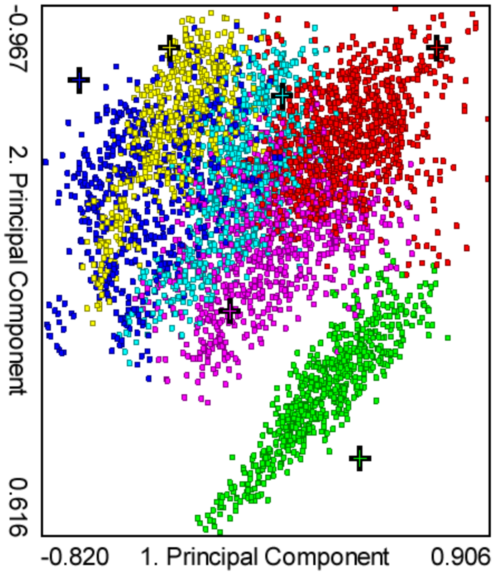

Figure 5: Repositioned cluster centers in principal components

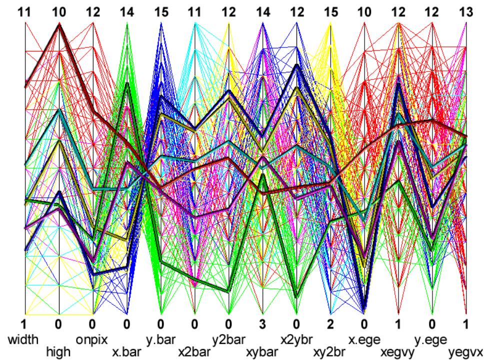

Figure 6: Repositioned cluster centers in parallel coordinates

This group finding process evaluated a similar solution than the initial clustering. Nevertheless the energy function of the k means clustering signalizes that a better result was found, because the sum of distances of the data items to their nearest cluster centers was reduced. The scatterplot in figure 7 reveals that the regions of overlap between the red, cyan and magenta cluster has been reduced, while the yellow and the blue cluster still interweave. To resolve this visual inseparability further principal components have to be taken into account. Only little changes can be observed in the parallel coordinates visualization (figure 8). As the clusters still show extreme values in single dimensions a subspace clustering can be considered, to separate the measurements of one letter from the remainder of the data.

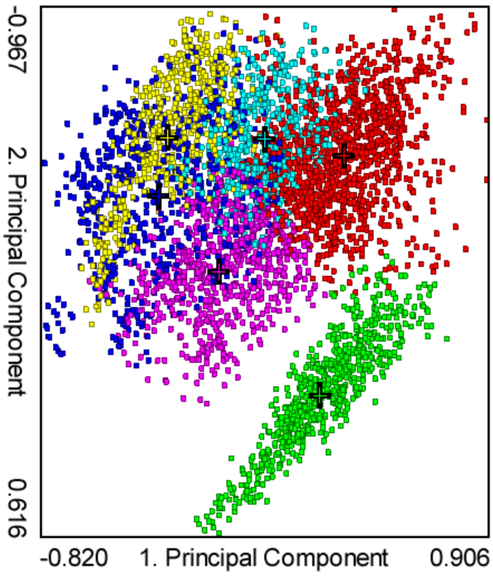

Figure 7: Reclustering result mapped on principal components

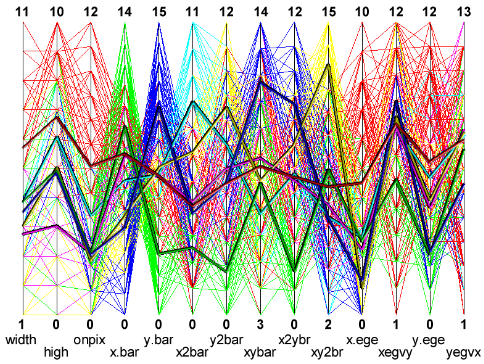

Figure 8: Reclustering result in parallel coordinates

6 Implementation

The implementation of the statistics library is aimed to operate on large data sets holding millions of data items. Therefore special attention was paid to process large arrays of data values as fast as possible. To achieve this goal an implementation in the language C++ was chosen, which provides efficient pointer operations. The work with C++ also demands a concept for a failure safe usage of the library. To accomplish this issue all routines return a bool variable indicating false, if the functionality could not be executed correctly. Furthermore the parameters that are passed to the methods apply to a scheme so that function calls of different procedures have a similar structure and thus are intuitive to use. The first category of parameters concerns variables into which the results are written. Afterwards data specific information like dimension values or mean vectors have to be set. The final class of parameters is concerned with the properties of the applied algorithm itself. Examples therefore are the number of clusters for a *k* means clustering, or the robustness factor for an MCD covariance matrix estimator.

To ease the integration of statistical routines into information visualization applications the definition of an adequate interface is crucial. Thus so called hooks of interaction have been realized. These special function calls enable the immediate recalculation of statistical facts like correlations and moments for subsets of the data items. This is important for applications where numerical summaries of selected data items are requested. Those summaries have to be updated if the selection changes or if details-on-demand actions are initiated. Besides these standard adaptations also task specific interface extensions had to be included. Based on the statistical routine and its visualization several interaction techniques can be specified. Consequently the implications of the user actions have to be translated into parameter settings for the computational algorithm in the statistical library to adapt its result. This was accomplished on the basis of the *k* means clustering. A typical visualization of a result of this partitioning procedure involves the representation of the data items in colors according to their cluster membership and the emphasis of the cluster centers. The latter can be used to manipulate the clustering result by repositioning its centers or selecting clusters to initiate merge and splitting operations. Those complex adaptations have to be covered by the hooks of interactions and reformulated into ordinary function calls to allow a reclustering based on the user's input.

As functionalities like the principal component analysis and the robust distance calculation require matrix inversion and determinant evaluation, implementations for these operations were integrated from the numerical recipes in C [20]. Also correlation computations, the realization of theoretic distributions and the Kolmogorov-Smirnov test build up on the fast and stable routines of this repository of basic numerical procedures. Because of the integration of robust methods, which require the evaluation of a value, having a specified position in a sample, an efficient routine that returns the *k* smallest data value of an array was realized [9]. Empirical tests prove that this method is faster than the application of a quick sort procedure.

7 Conclusions

While the most applications that consider techniques of both information visualization and statistics, concentrate on presentation and exploration of the results of computational routines, the possible benefit of an interactive collaboration of these fields is higher.

Future work has to concentrate on the translation of interactions performed in a visual data mining view into parameter settings for statistical algorithms. Without this contribution an efficient combination of the strengths of computational capabilities with the experience and the knowledge of the user is not possible. Also the interactive information exchange between statistics and user actions has to be accomplished. Numerical summaries, that are immediately updated if selections are changed or details-on-demand operations are executed, are necessary to validate the percieved patterns in the visualizations.

Furthermore the concept of robustness, which is well known in the field of statistics, has to be incorporated into the visual data mining workflow. This is a crucial task, because outlying data items can significantly distort the results of statistical routines as well as the visual impression provided by visualizations, which build up on the outcome of these algorithms.

8 Acknowledgement

This work has been realized at the VRVis Research Center in Vienna, Austria (http://www.VRVis.at/), which is funded by the Austrian research program called Kplus, to support the basic resarch on visualization (http://www.VRVis.at/vis/).

Special thanks are given to Helwig Hauser, who gave me the opportunity to work on the combination of statistical routines and information visualization. His support and his vision how both sciences could collaborate made this work possible.

I also want to thank Peter Filzmoser for his input and help concerning statistical routines and their integration in visualization applications. His ideas for the improvement of data exploration by visual techniques was a crucial contribution to increase the quality of this work.

Additional thanks go to Harald Piringer, whose dedication for improving implementations and whose expert knowledge concerning the work with large datasets helped to realize the assembly of statistical routines in a C++ library.

References

[1] GGobi - Data visualization system (http://www.ggobi.org/).

[2] Gretar Tryggvason 3-d fluid flow data, (XmdV data archive: http://davis.wpi.edu/xmdv/datasets/uvw.html).

[3] GTK+ The GIMP Toolkit (http://www.gtk.org/).

[4] Miner3D (http://www.miner3d.com/).

[5] SpotFire Decisionsite (http://www.spotfire.com/products).

[6] James Christian Bezdek. *Fuzzy mathematics in pattern classification*. PhD thesis, Cornell University, 1973.

[7] Markus M. Breunig, Hans peter Kriegel, Raymond T. Ng, and Jrg Sander. LOF: Identifying density-based local outliers, March 01 2000.

[8] William S. Cleveland. *The Elements of Graphing Data*. Hobart Press, 1994.

[9] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 2001. Second Edition.

[10] Peter W. Frey and David J. Slate. Letter recognition using holland-style adaptive classifiers. *Machine Learning*, 6:161, 1991.

[11] Diansheng Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2(4):232–246, 2003.

[12] J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

[13] Alexander Hinnenburg, Daniel A. Keim, and Markus Wawryniuk. HD-eye: Visual mining of high-dimensional data. *IEEE Computer Graphics and Applications*, 19(5):22–31, September 1999.

[14] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *VIS '90: Proceedings of the 1st conference on Visualization '90*, pages 361–378. IEEE Computer Society Press, 1990.

[15] J. E. Jackson. *A User's Guide to Principal Components*. Wiley, New York, 1991.

[16] Keim and Hinneburg. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In *VLDB: International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers, 1999.

[17] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB '98)*, pages 392–403, East Sussex - San Francisco, August 1998. Morgan Kaufmann.

[18] Teuvo Kohonen. The self-organizing map. *Proc. IEEE*, 78(9):1464–1480, September 1990.

[19] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Piblications, Beverly Hills, CA, 1978.

[20] W. H. Press et al. Numerical recipes in C (second edition). *Cambridge University Press*, 1992.

[21] Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator, 1998.

[22] SGI. OpenGL - reference manual, the official reference documantation for openGL, release 1. *Addison Wesley, ISBN 0-201-63276-4*, 1992.

[23] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *VL*, pages 336–343, 1996.

[24] Jing Yang, Matthew O. Ward, Elke A. Rundensteiner, and Shiping Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. pages 19–28.

[25] Jing Yang, Matthew O. Ward, Elke A. Rundensteiner, and Anilkumar Patro. Interring: a visual interface for navigating and manipulating hierarchies. *Information Visualization*, 2(1):16–30, 2003.