A Multidimensional Brush for Scatterplot Data Analytics

Michaël Aupetit* Qatar Computing Research Institute

Nicolas Heulot† IRT SystemX

Jean-Daniel Fekete‡ AVIZ INRIA

ABSTRACT

Brushing is a fundamental interaction for visual analytics. A brush is usually defined as a closed region of the screen used to select data items and to highlight them in the current view and other linked views. Scatterplots are also standard ways to visualize values for two variables of a set of multidimensional data. We propose a technique to brush and interactively cluster multidimensional data navigating through a single of their scatterplot projection.

Keywords: Visual analytics, brushing, scatterplot, multidimensional data, axis-parallel projection.

Index Terms: Visual analytics; Cluster analysis; Statistical graphics

1 INTRODUCTION

Brushing is a standard interaction technique for visual analytics. A brush is usually a closed region of the display used to select data items and to highlight them in the current view and other linked views. The reason for brushing data can be the focus on specific visual patterns like outliers or clusters in one view that the user wants to link to other views for further analysis. It is also a way to select data to be filtered out or to be highlighted permanently to keep track of detected patterns for future interactions. Clusters are important patterns in Exploratory Data Analysis. Data within a cluster are more similar to each other than to other data clusters.

Two-dimensional (2D) scatterplots are also a standard way for data visualization where each data is represented as a point in a 2-dimensional Cartesian space. When data are multidimensional (MD), 2D scatterplots can show an orthogonal projection of the data onto the plane formed by two of the data variables. However, this projection hides the MD cluster structure by possibly overlapping several distinct MD clusters in a single one. Even Scatter Plot Matrices (SPLOM) are unable to show this cluster structure for three possible reasons: the clusters are not pairwise linearly separable (nested clusters); the clusters are pairwise linearly separable but not by a hyperplane orthogonal to any one or two of the MD variables (Blue/red and red/green points in Fig. 1); or the clusters are pairwise linearly separable by such a hyperplane but this is hidden by another cluster lying in between in any of the 2D scatterplots (Blue/green points separability hidden by the red points in Fig. 1b).

In this work we focus on an interactive clustering task (a standard high-level analytic task in Exploratory Data Analysis) and define an MD brush which enables the user to solve these issues, exploring similarities between MD data through their 2D scatterplot, and to keep track of this exploration by coloring the MD clusters found.

* email: maupetit@qf.org.qa
† email: nicolas.heulot@gmail.com
‡ email: Jean-Daniel.Fekete@inria.fr

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

2 MULTIDIMENSIONAL BRUSHING

We consider a set P of D-dimensional data $x=(x_1,...x_D)$ as vectors in the data space $E=IR^D$, their orthogonal projection as a scatterplot of points (x_a,x_b) (a and b in $\{1,...,D\}$) in the 2D plane E_{ab} and we define a brush B_S as a ball in S subspace of E with Euclidean radius r_S and center v_S. In the sequel every distances and radii are Euclidean.

2.1 Brushing in 2D space

A standard brushing $B_{ab}(P)$ in the scatterplot E_{ab} consists in selecting points from P whose distance to $v_{ab}=(v_a,v_b)$ is lower or equal to r_{ab}. The points selected are assigned a color or shape contrasting to non-selected points. The radius r_{ab} of this 2D-brush can be tuned with a slider or the mouse wheel, while its center position v_{ab} can be assigned to the mouse pointer position. This brush selects data points in the subspace $Q_{ab} = B_{ab} \times E_{\backslash a \backslash b}$. The brush is local in E_{ab} but not in $E_{\backslash a \backslash b}$ ($E = E_{ab} \times E_{\backslash a \backslash b}$) so it ignores any cluster structure in the $E_{\backslash a \backslash b}$ subspace as illustrated in Fig. 1.

2.2 Brushing in MD space

The multidimensional brush (MD-brush) is obtained by setting $S=E$, the MD-brush is then a D-dimension ball B_E with radius r_E centered at some point v in E, that can be visualized as a disc B_{ab} with radius r_{ab} and center $v_{ab}=(v_a,v_b)$ in the scatterplot E_{ab}. This B_{ab} disc defines a Magic Lens [1]: the appearance of the data points lying within this disc is altered based on the MD-brush selection process.

MD-brush selection: Data whose MD distance to v_E is lower or equal to r_E are lying within the MD-brush. Among these points, only the ones who also lie within the 2D magic lens are selected. We highlight these selected points by changing their shape (Fig. 2.). This highlighting in the scatterplot is linked to the current MD-brush selection and changes instantly with it.

MD-brush positioning: The mouse pointer position in E_{ab} corresponds to a whole D-2 subspace of E, so it cannot be used in general to set the MD position of the MD-brush center v. Therefore we propose to set v to the MD position of the nearest 2D data point to the mouse pointer.

Magic lens radius r_{ab} tuning: The radius of the magic lens can be tuned using the mouse wheel. The smaller the radius, the more local the analysis of the MD cluster structure in the neighborhood of the MD-brush center v but the lower the number of points selected to get statistically relevant outcomes.

MD-brush radius r_E tuning: The radius of the MD-brush is not straight-forward to tune. Indeed, while it would seem natural to set r_E equal to r_{ab}, the probability for B_E to be empty while B_{ab} is not would increase with D. This is an effect of the curse of dimensionality [2]. If distances between data in E_{ab} scale with unity on average, then MD distances between data in E scale with the square root of D. Moreover if we consider a D-variate unit variance Normal distribution of data points centered at v, the distances of the data points to v follow a Chi distribution, so the smallest r_E so that B_E captures all the points lying within B_{ab} may be far larger than r_{ab}. We propose to support the user by visualizing as a bar graph, the distribution of the MD distances of any data points in B_{ab} to the center v of B_E together with the Chi distribution of the distances of points drawn from an M-variate

Normal distribution centered at v with a diagonal covariance matrix Σ_E whose non-zeros elements are all equal to $\sigma^2 = (r_{ab} / 2.45)^2$. The radius of the 95% quantile circle of a bivariate unit variance Normal is equal to 2.45. So if MD data would come from a single M-variate Normal centered at v with variance Σ_E, then their 2D projection in E_{ab} would be also a Normal with center v_{ab} and variance σ^2 so that the Magic Lens B_{ab} would contain 95% of the data. Thus this setting assumes the user selects a 2D cluster in B_{ab} and sees how the distances would be distributed if all of the data within the magic lens would represent the 95% core mass of a single M-variate Normal cluster. This allows the user to tune r_E so the empirical MD distance distribution within B_E is similar to the theoretical Chi distribution thus defined. If the empirical distribution appears to be denser than the theoretical one for smaller MD distances, that means points are even more concentrated or lie within a lower dimensional subspace clustered around v. We draw a vertical line indicating the 95^{th} percentile of the Chi distribution which can be used as a default setting for r_E (Fig. 2 and 3).

MD-brush clustering: A control key enables the user to permanently color the selected points to keep track of MD clusters found. The interactive clustering process we propose is in the spirit of density-based automatic clustering approaches [3]. The user starts positioning the Magic Lens at some point in the scatterplot, then tunes the MD-brush radius r_E to get a near Normal-cluster-like MD distance distribution, then assigns the selected points to the current cluster and keeps up exploring the border of the current MD cluster point by point to enlarge or retract it. Notice that focusing the next MD-brush on a currently selected point in the 2D scatterplot corresponds to navigate continuously from a data point to its neighbors in the MD space E.

3 Analytic Use Case

The data used are letters {A, B, C, D, E} from the ISOLET dataset of the UCI repository. These letters were pronounced twice by 150 English speakers, 617 features were extracted from the signals. We kept the first 30 speakers for each letter so we got 300 instances of 617-dimension data that we further reduce to keep the 10 leading principal components using PCA. We attempted to extract manually the 5 clusters using the MD-brush through the scatterplot of the first two principal components. The interactive process of MD-brushing is demonstrated in Fig. 2, 3 and 4. At the end we got 3 clusters, one of which appeared to contains letters {B, D, E} (Fig. 4). We checked (Table 1) that {B, D, E} clusters are not mutually linearly separable in the 10D space, while they were each linearly separable with {A, C} clusters, supporting our findings with MD-brush.

4 Discussion

We showed that an MD-brush can help to explore MD data through a single 2D scatter plot visualization, and recover MD cluster structures that other standard scatterplot-based visualization like SPLOM even equipped with brush and link, would fail to reveal. MD-brush relies only on the MD distances between the data points, so it could be used with other MDS-like projection but this has not been studied yet. Moreover, this work is to be complemented with a user-study to test how intuitive it is for non-expert users to recover MD clusters within various datasets and how the design could be improved for instance to display the distances' histogram within the scatterplot. At last, a future work would be to study how MD-brush could be extended combined with SPLOM, to select the space E to be explored for subspace clustering.

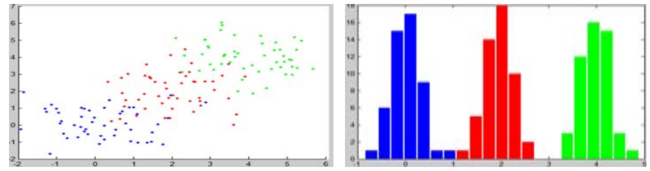

Figure 1: overlapping scatterplots along two variables of data drawn from three separated 10D unit variance Normal centered at (0,…,0); (2,…,2) and (4,…,4), and histogram of the data projected onto the [(0,…,0);(1,…,1)] line. The user has to discover the clusters (colors).

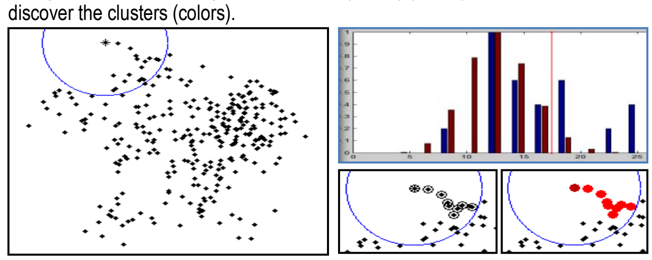

Figure 2: The first 2 principal components of the 10D ISOLET data (black dots). The blue circle is the Magic Lens associated to the MD-brush. The histogram shows the empirical distance distribution (blue) to the brush's center (star), and the theoretical distance distribution (brown) for 10D data within this circle (Chi distribution). Setting the radius r_E at the red line value provides the MD-brush selection (circled dots) then assigned to red cluster (dark red for focus point).

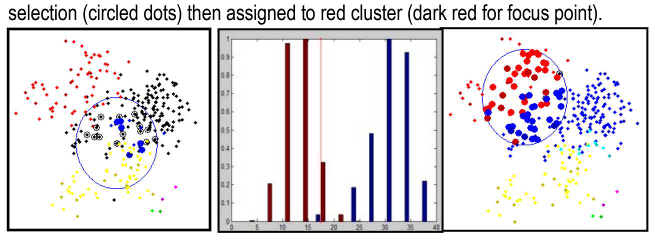

Figure 3: After some MD-brush clustering (left), the circled black dots and the blue spots are within the current MD-brush while yellow dots are not, therefore the circled black dots will be added to the blue cluster. Later (right), a case where no clear cluster appears (no MD distance on the left of the red line), setting r_E to 40 shows that the center lies at the border of blue and red clusters. It will be assigned arbitrarily to the blue one.

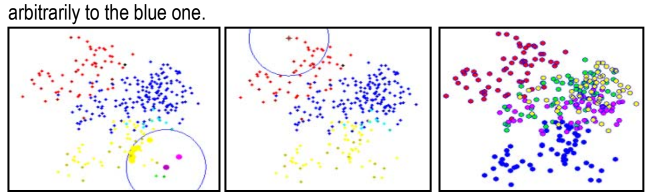

Figure 4: Coming back to the small magenta cluster (left) it appears to be within the MD-brush with many data from the yellow cluster (yellow spots). The same is true for the small green cluster (not shown). Both will be finally assigned to the yellow one. The final MD-brush clustering result (center) and the true classes (right, with letters {**A**, **B**, **C**, **D**, **E**} encoded as red, green, blue, magenta, and yellow colors respectively). The yellow MD-brush cluster is very similar to the true blue class, the red one to the true red class, and the dark blue to the union of magenta, yellow and green true classes.

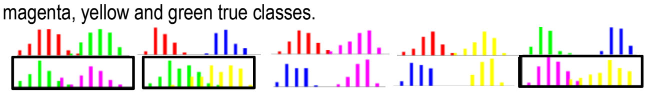

Table 1: Classes pairwise linear separability in the 10D PCA space of the ISOLET data. Pairs {**B**, **D**}, {**B**, **E**} and {**D**, **E**} are not linearly separable (black frame).

References

[1] C. Tominski, S. Gladisch, U. Kister, R. Dachselt, and H. Schumann. A Survey on Interactive Lenses in Visualization. EuroVis State-of-the-Art Reports, Eurographics Association, 2014.

[2] D. François, V. Wertz, and M. Verleysen. About the locality of kernels in high-dimensional spaces. Proc. of ASMDA, pp. 238-245. Brest (France), 17-19 May 2005.

[3] H.-P.Kriegel, P.Kröger, J.Sander, A.Zimek. Density-based Clustering. Data Mining & Knowledge Discovery 1(3): 231-240, 2011.