UvA-DARE (Digital Academic Repository)

Visualizing Multi-Dimensional Decision Boundaries in 2D

Migut, M.A.; Worring, M.; Veenman, C.J.

DOI

[10.1007/s10618-013-0342-x](https://doi.org/10.1007/s10618-013-0342-x)

Publication date

2015

Document Version

Final published version

Published in

Data Mining and Knowledge Discovery

[Link to publication](#)

Citation for published version (APA):

Migut, M. A., Worring, M., & Veenman, C. J. (2015). Visualizing Multi-Dimensional Decision Boundaries in 2D. *Data Mining and Knowledge Discovery*, 29(1), 273-295.
<https://doi.org/10.1007/s10618-013-0342-x>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Visualizing multi-dimensional decision boundaries in 2D

M. A. Migut · M. Worring · C. J. Veenman

Received: 21 February 2012 / Accepted: 12 December 2013 / Published online: 28 December 2013
© The Author(s) 2013

Abstract In many applications experts need to make decisions based on the analysis of multi-dimensional data. Various classification models can support the decision making process. To obtain an intuitive understanding of the classification model, interactive visualizations are essential. We argue that this is best done by a series of interactive 2D scatterplots. In this paper, we define a set of characteristics of the multi-dimensional classification model that have to be visually represented in those scatterplots. Our proposed method presents those characteristics in a uniform manner for both linear and non-linear classification methods. We combine a visualization of a Voronoi based representation of multi-dimensional decision boundaries with visualization of the distances of the data elements to these boundaries. To allow the developer of the model to refine the threshold of the classification model and instantly observe the results, we use interactive decision point selection on a performance curve. Finally, we show how the combination of those techniques allows exploration of multi-dimensional decision boundaries in 2D.

Keywords Interactive data mining · Knowledge discovery · Decision boundary visualization · Multi-dimensional space · Classification · Visual analytics

Responsible editor: Eamonn Keogh.

M. A. Migut (✉) · M. Worring
Intelligent System Lab Amsterdam, University of Amsterdam,
Amsterdam, The Netherlands
e-mail: mmigut@gmail.com

M. Worring
e-mail: M.Worring@uva.nl

C. J. Veenman
Digital Technology and Biometrics Department, NFI,
The Hague, The Netherlands
e-mail: c.veenman@nfi.minjus.nl

1 Introduction

In many domains, experts have to make decisions based on the analysis of multi-dimensional data. The core element of the decision making process is an accurate and transparent classification model. For the validation of the model and for deepening the insight into the domain and its underlying processes, intuitive means to assess the characteristics of the classifier are important.

To assess the classifier it is favorable to obtain a visual comprehension of its most important characteristics. Many visualization techniques are in use to present the results of the classifier, such as ROC curves or Precision and Recall graphs (Duda et al. 2000). These are very useful techniques for visualizing, organizing, and selecting classifiers based on their performance (Provost and Fawcett 1997). Performance alone, however, is not sufficient to understand a classifier. The resulting graphs are an aggregation of classification results of individual data elements, so the contribution of each data element to the result remains unclear. Consequently, it is difficult to assess which elements are easily classified, which are more complex to handle, and what mistakes are made. This is so difficult because the results stem from an intricate interplay between the characteristics of the data, the classification model, and its parameter settings. For selecting the most suitable classifier we need to go beyond performance as the only measure. Furthermore, the best overall performance is often not the optimal solution in applications where risk or profit are assessed, e.g. in disaster management, finance, security, and medicine (Keim et al. 2008; Thomas and Cook 2005). In the medical field, for example, diagnosing specific disorders is a common task performed by a medical expert. The consequences of making mistakes, either diagnosing the healthy patients or not diagnosing the ill patients, may be fatal. Obviously the expert has a difficult task to understand the multi-dimensional data, make decisions based on it and moreover foresee the consequences of those decisions. Crucial in the decision making process is determining the cost of the different types of mistakes made by the classification model. From a holistic point of view, the differences in the amount of mistakes made by different classification models for the same dataset are often statistically insignificant. However, these classification models do not make the same mistakes. Therefore, assuring that critical cases are handled appropriately, and finding the balance between different types of mistakes are important.

One of the most informative characteristics of the classification model and its relation to the data is the decision boundary. It is the boundary which determines the areas in space where the classes are residing. It also provides a reference for determining classification difficulty. Elements close to the boundary are the ones which are difficult to classify, while the more distant ones have a higher certainty of class membership. Being able to visualize the decision boundary is a great aid in decision making. Decision boundaries can easily be visualized for 2D and 3D datasets (Duda et al. 2000). Transforming a boundary in 4 or more dimensions to a representation in lower dimensions, that can be displayed and understood by the experts is difficult. The most important cause is the inherent loss of information when mapping from high- to low-dimensional space. Properly defining the core characteristics of the multi-dimensional decision boundary that should be represented in the low dimensional representation is crucial.

Several attempts have been made to visualize decision boundaries for multi-dimensional data (Caragea et al. 2001; Hamel 2006; Poulet 2008; Migut and Worring 2012). The first two methods are specific to limited types of classifiers and hence can not be used to compare different methods. The method in (Poulet 2008) can be applied to different classifiers, however it does not allow the user to relate the visualization of the decision boundary to the data elements in terms that are meaningful for domain experts. The method in (Migut and Worring 2012) does not allow users to analyze the data elements in relation to the decision boundary as the distances to the boundary are not visually expressed. None of those approaches on their own allow the expert to examine different classifiers in terms of the decision boundary and costs of classification in terms of misclassified examples.

The aim of this paper is to expand on the previous studies (Migut and Worring 2012; Poulet 2008) to find a generic approach to visually analyze a decision boundary of a multi-dimensional classifier in 2D. To this end, we formalize the problem by providing a set of decision boundary characteristics that the 2D visualization should represent. Moreover, we formalize the approach taken by (Migut and Worring 2012) and by combining it with methods proposed by (Poulet 2008) we provide an expert with a visualization of the decision boundary that expresses all the important characteristics of the classifier and allows the analysis of classification results. We interactively couple the data visualization, the decision boundary visualization, classifier performance visualizations and the distance to the decision boundary. The integrated solution provides an expert with the possibility to visually explore the classifier and the costs of classification, as well as visually compare different classifiers for all data dimensions.

The paper is organized as follows. First, we formally state the problem, define a set of tasks that require a visual representation of the decision boundary, and propose a set of classifier's characteristics that should be visually expressed in 2D. We also describe in detail why the visualization of a decision boundary in 2D is so challenging. The subsequent section presents a review of several attempts to visualize multi-dimensional decision boundaries. We pin-point the shortcomings and potential of the existing techniques. From there, we show how to interactively integrate several visualization techniques that contribute to a solution satisfying our requirements. Finally, we demonstrate the approach using five datasets illustrating that the proposed methodology is suitable for exploring multi-dimensional classifiers.

2 Problem analysis

In this section we formalize our problem and propose a set of decision boundary characteristics a successful visualization should represent.

2.1 Characteristics of the classifier

We assume a training dataset with k objects represented by feature vectors with numerical data in an n -dimensional metric space. For now, we restrict ourselves to two-class problems. In Sect. 5.3, we indicate how to transfer the techniques to multi-class problems. Also, the primary feature spaces we target are those having meaning to the

expert, so n is small. In the first phase, a classifier is trained on the dataset, resulting in a decision boundary in the n -dimensional space. Based on the boundary new objects are classified. An object classified as positive is defined as true positive (TP) if the actual label is also positive and is called false positive (FP) if the actual label is negative. In a similar way, an object classified as negative is called true negative (TN) if the actual label is negative and false negative (FN) if the actual label is positive.

The problem at hand is how to visualize the n -dimensional decision boundary to support the expert's decision making process. To that end, let us first look at the main tasks that the expert has to perform. We summarize them as follows:

- [T1] Task 1: Analyze which of the n dimensions are most important
- [T2] Task 2: Analyze and compare how different classifiers separate the data
- [T3] Task 3: Analyze the relation between boundary and data elements
- [T4] Task 4: Analyze and compare classification costs

These tasks have several implications that make the visualization of the n -dimensional decision boundary a challenging task. In particular, there are four important characteristics of the classifier that a visualization of the decision boundary should capture:

1. **Separation:** the visualization must be in agreement with the actual classification. All objects assigned to a positive class by the classifier (TP and FP) must be visually differentiated from the members of the negative class (TN and FN). On the level of the individual data objects, the visualization must represent whether there is a decision boundary between each pair of objects in the multi-dimensional space.
2. **Direction:** for two arbitrary data objects in the visualization, the representation of the decision boundary must unambiguously show on which side of the decision boundary each object is located in the multi-dimensional space.
3. **Distance:** for each visualized data object the distance to the decision boundary in multi-dimensional space should be represented.
4. **Performance:** the classification result of the current decision boundary should be represented.

To compare different classifiers, the visualization technique must consistently represent the characteristics independently of the classification method used.

2.2 Data visualization technique

The data visualization technique must be chosen such that it allows the visualization of these characteristics in a way that supports the conceptual framework of the experts using the system. A taxonomy of candidate multidimensional visualizations is given by (Keim 2002). The categories listed include standard 2D/3D displays, geometrically transformed displays, iconic displays, dense pixel displays, and stacked displays. Different techniques serve different purposes. Since experts understand their data best in the original feature values, we consider here only techniques that support this. Interactive visualizations of multi-dimensional datasets which represent the features explicitly are e.g. scatterplots, heatmaps, parallel coordinates and parallel sets (Bendix et al. 2005).

Out of the elements in the taxonomy scatterplots are the most used building blocks in statistical graphics and data visualization (Cleveland and McGill 1988). Multidimensional visualization tools that feature scatterplots, such as Spotfire, Tableau/Polaris (Stolte et al. 2002), GGobi (Swayne et al. 2003), and XmdvTool (Ward 1994) typically allow mapping of data dimensions also to graphical properties such as point color, shape, and size. In a 2D scatterplot data elements are drawn as points in the Cartesian space defined by two graphical axes defined by the real attributes values (in domain space). Thus they accommodate the conceptual framework of the expert user. Moreover their familiarity among users favor their use for the purpose of this paper. However, the number of dimensions that a single scatterplot can visualize is considerably less than what is found in realistic datasets. Therefore, several scatterplots should be visualized for all combinations of the dimensions. This is usually done using a scatterplot matrix (Cleveland and McGill 1988), interactive scatterplot matrix navigation (Elmqvist et al. 2008) or simply using the changeable axis, where the user interactively selects the attributes to be plotted. In all of the above methods the variables are plotted against each other preserving the meaning of their values.

In case the number of dimensions is too high to show and explore all dimensions, the amount of dimensions have to be reduced. In machine learning, dimension reduction can be divided into feature extraction and feature selection. Feature extraction transforms data from a high-dimensional space into a space of fewer dimensions. The transformation can either be linear or non-linear. In both cases the original dimensions that are known to the user are lost. Feature selection techniques reduce the number of dimensions by finding a subset of original features that are most relevant for use in the model construction. As opposed to feature extraction, after feature selection the original features can be visualized, therefore we favor this approach. Consequently, the decision boundary in multi-dimensional space has to be visualized in such a 2D setting.

2.3 Axes parallel projections

The decision boundaries for multi-dimensional classifiers are either planes/hyperplanes for linear classifiers or can exhibit complex shapes for non-linear classifiers. For two dimensions at the time, the multi-dimensional data can easily be projected into 2D space. The resulting projections do not necessarily separate elements belonging to different classes as imposed by the multi-dimensional classifier. Only for the multi-dimensional linear decision boundary that is perpendicular to the projection plane, the separating information will be preserved. For linear boundaries, as well as non-linear boundaries, the straightforward projection captures neither the **Separation** nor the **Direction** characteristics of the multi-dimensional classifier. **Distance** is also not represented. This follows from the observation that distances in the 2D data projection do not represent the actual distances in the multi-dimensional space. The distances of data elements to the projected boundary would, therefore, also not be preserved.

Therefore, the straightforward projection of the multi-dimensional boundaries to 2D is not the answer to our problem. We will look for a methodology to represent the multi-dimensional decision boundary, that will allow to see how the classifier separates the data in the original multi-dimensional space, when the data is projected into 2D.

3 Related work

Many recent papers, especially in the Visual Analytics community, focus on different approaches to combine visualization and automated data analysis (Malik et al. 2012; Hoferlin et al. 2012; Brown et al. 2012; Endert et al. 2011; Jeong et al. 2009), especially in the context of supervised machine learning (Choo et al. 2010; Heimerl et al. 2012). Only few attempts have been made to visualize decision boundaries for multi-dimensional data (Poulet 2008; Caragea et al. 2001; Hamel 2006). We summarize those techniques and analyze which characteristics of the decision boundary they capture.

In (Caragea et al. 2001) authors visualize the Support Vector Machine classifier (SVM) using a projection-based tour method. The authors show visualizations of histograms of the data predicted class, visualization of the data and the support vectors in 2D projections and weighting the plane coordinates to choose the most important features for the classification. The methods are all applicable to SVM only.

An interesting approach to decision boundary visualization is proposed in (Poulet 2008). The authors display the histograms of the distance to the boundary distribution of correctly classified examples and misclassified examples. Those histograms are linked to a set of scatterplots or parallel coordinates plots. The bins of the histogram can be selected and the points on the scatterplot with corresponding distances to the multi-dimensional boundary are consequently highlighted. The authors claim that those highlighted elements are showing the separating boundary on the scatterplots. However, if for a certain 2D projection, the elements close to the decision boundary are scattered all over the plot, it is no longer possible to understand how the classifiers separates the data. This means that it is not possible to directly assess whether there is a decision boundary between two arbitrary points in the visualization. This method does offer a good estimation of the quality of the boundary for SVM, and can also be applied to other classifiers like decision trees or regression lines. However, linking the histogram of the distances to the decision boundary with the corresponding points on the scatterplot is not enough to give an insight into how the classifier separates the data.

The method in (Hamel 2006) uses self-organizing maps (SOM) to visualize results of SVM. SOM's are also used to visualize decision boundaries in (Yan and Xu 2008). Yan proposes two algorithms, one to obtain data points on decision boundaries and a second one to visualize decision boundaries on SOM maps. The decision boundaries are not visualized in the original feature space (i.e. domain space).

We conclude that even though the described techniques to visualize decision boundaries are very interesting, none of them alone can be used to show all the characteristics of the decision boundary and the costs of classification for different classifiers.

4 Voronoi based decision boundary visualization

In this section we describe and formalize the Voronoi-based representation of the decision boundary, as used in (Migut and Worring 2012). In order to capture the characteristics of the classifier to represent the decision boundary, we extend that technique with the visualization of the histogram of distances, as used by (Poulet 2008).

4.1 Separation

Let us consider two different elements in the dataset, a labeled as positive by our classifier C and b labeled as negative. The **Separation** characteristic of the classifier implies that in the visual representation of classifier C the boundary must lie somewhere between point a and b . If we assume it to be locally linear it would yield a half plane containing a and not b . Due to the limits of the 2D space compared to the high dimensional it is virtually impossible to preserve the actual distances. So not being able to preserve distances we might as well put it midway the two elements. If performed for all the elements in the dataset, this yields a Voronoi tessellation of the space. Therefore, we use a Voronoi tessellation to represent the decision boundary in a 2D scatterplot.

A Voronoi diagram (Fortune 1987; Aurenhammer 1991; Duda et al. 2000) can be described as follows. Given a set of points (referred to as nodes), a Voronoi diagram is a partition of space into regions, within which all points are closer to some particular node than to any other node, see Fig. 1. Formally, if P denotes a set of k -points, then for two distinct points $(p, q) \in P$ the separator separates all points of the plane closer to p from those closer to q :

$$sep(p, q) = \{x \in R^2 | \delta(x, p) \leq \delta(x, q)\},$$

where δ denotes the Euclidean distance function. The region $V(p)$ being the Voronoi cell corresponding to a point $p \in P$, encloses part of a plane, where $sep(p, q)$ holds:

$$V(p) = \bigcap_{q \in P - p} sep(p, q).$$

Fig. 1 **a** Voronoi diagram for a 2-dimensional dataset of two Gaussian distributed classes together with the approximated decision boundary following the Voronoi cells' boundaries (*thick solid line*). The approximation follows the labels as imposed by the classifier (linear support vector machine) and therefore does not violate the actual classifier visualized with a *dashed line*. **b** ROC curve with the current classifier's trade-off visualized as an operating point on the curve

Fig. 2 Decision boundary for a 10-dimensional dataset of two Gaussian distributed classes for the SVM classifier: **a** Projection of the boundary to a 2D space spanned by 2 features which are separating the data well, **b** Idem, but now for 2 features which yield a highly fragmented boundary

Two Voronoi regions that share a boundary are called Voronoi neighbors. We apply the Voronoi diagram to each combination of two dimensions projected into 2D space for a dataset labeled by the multi-dimensional classifier C . All the data objects are used as nodes to create the Voronoi diagram. The boundaries of the Voronoi regions corresponding to neighbors belonging to different classes (according to the labels assigned by the classifier) form the decision boundary. Such a representation of class separation for two given features is a piecewise linear approximation of the actual decision boundary, as imposed by the multi-dimensional classifier, see Fig. 1.

Visualization of the 2D combinations of the dimensions used by the classifier results in a series of scatterplots, where each pair of features might exhibit correlation, good class separation or other interesting characteristics. For the features that separate the data well, the approximated decision boundary ‘disconnects’ the classes well, resulting in two clusters of data. For the features that do not separate the data well, we observe a high fragmentation of the classes. Figure 2 illustrates our approach.

4.2 Direction

According to our definition, for two arbitrary data objects in the visualization, the representation of the decision boundary must unambiguously show on which side of the decision boundary each object is located in the multi-dimensional space. The visualized elements that belong to the different classes according to the classifier must therefore have different visual appearances. We call this characteristic (on which side of the boundary the elements are located): the direction. Since the color represents the original labels, we assign different shapes to distinguish between the predicted labels. The shaped-based difference in appearance is easily noticeable if there are few data elements plotted. However, when the scatterplot becomes cluttered, difference between classes is less apparent. In order to emphasize on which side of the boundary the elements are located, we therefore color the Voronoi regions belonging to one class. We argue, that for complex boundaries, a region based visualization makes it easier to

Fig. 3 **a** The histogram of the distances of the TP, TN, FP, FN to the decision boundary, with the highlighted bin of the closest TP to the boundary, as proposed in (Poulet 2008) **b** The True Positives with the closest distance to the decision boundary highlighted. We see which elements are the closest but because they are scattered all over the plane, they do not indicate how the decision boundary is related to other data elements, **c** Again, the TP with the closest distance to the boundary are highlighted. Due to the Voronoi based representation of the decision boundary it is immediately visible how the boundary divides that 2D representation of the data

distinguish on which side of the multi-dimensional decision boundary each element is located and therefore to comprehend to which class each data instance belongs. Figure 3c illustrates how color can be used to show the direction of the representation of the Voronoi decision boundary.

4.3 Distance

As stated in the previous subsection, the consequence of using scatterplots of multi-dimensional data is that the distances between data points from the original domain space are not preserved. Therefore, the distances between the data elements and the visualized decision boundary can also not be preserved. Using the Voronoi-based approach the distances between the actual objects and the decision boundary are indeed not preserved, but class membership is (i.e., separation + direction). In fact, the Voronoi based approximation of the decision boundary indicates precisely that in the multi-dimensional space there is a decision boundary between each two data instances located on the different side of the boundary representation. By construction, the piecewise linear representation lies exactly halfway the two points. This distance has no meaning in terms of the actual distance between the objects and the decision boundary in the original space. The position of the decision boundary could be optimized locally, between the two points that are in direct neighborhood of the linear piece of the boundary. However, the distance has also no meaning when considering the ordering of the data elements in any direction from the decision boundary. Therefore, the distances could not be optimized globally. This means we need another visual representation to show which of the elements are closer to the decision boundary.

To visually indicate the distances between the data points and the decision boundary, we exploit the histogram of the distances to the boundary (Poulet 2008). The histogram is divided into four regions. On the positive side of the X-axis the distances to the

elements with positive original label are visualized. The negative side of the X-axis is reserved for the data elements with negative original label. The positive part of the Y-axis is reserved for the elements that are correctly classified by the classifier and the negative side of the Y-axis is for the distances to misclassified examples. Figure 3a shows the four quarters. A histogram of distances allows making the difference in the distance to the multi-dimensional decision boundary visually distinct adding to the “completeness” of decision boundary visualization. The class-membership and actual distances to the boundary can be explored thorough interaction.

4.4 Performance: interactive trade-off inspection

In this section we show how to interactively couple performance trade-off visualizations with decision boundary visualizations, as proposed in Migut and Worring (2012).

In general performance curves such as Precision and Recall graphs or ROC curves capture the ranking performance of the binary classifier, as its discrimination threshold is varied. The Receiver Operating Characteristics (ROC) curve often used in medicine visualizes the trade-offs between hit rate and false alarm rate (McClish 1989). The Precision and Recall curve often used in information retrieval depicts the trade off between the fraction of retrieved documents relevant to the search and the fraction of the documents relevant to the query successfully retrieved. What these curves have in common is that they give a balance between two competing and inversely related measures. In applications where delicate decisions have to be made, this balance is subtle and complex as it can have dramatic consequences. Therefore, the performance curve should depict the trade-off between the classifier’s errors for one or both classes. As an example for this paper, we use the curve that represents the trade-off between the False Positive (FPr) rate and False Negative (FNr) rate. On the performance graph FPr is plotted on the X axis and FNr is plotted on the Y axis. These statistics vary with a threshold on the classifier’s continuous outputs. The trade-off of the current classifier is visualized by means of an operating point on the curve.

The relationship between the operating point on the ROC curve that corresponds with the decision boundary visualized using Voronoi tessellation can be formally described as follows. Let $V(p)$ be the Voronoi cell corresponding to a point p . For a set of points P we define $V(P) = \bigcup_{p \in P} V(p)$. For classifier C , let C_t be the decision boundary in n -dimensional space defined by the current operating point t . Further let $d(p, C_t)$ be a signed measure indicating how far element p is from the boundary where the sign of d is positive if p is classified to the positive class and negative otherwise. Given the set P of classified elements, the set of points classified to the positive class for the current operating point (P_t^+) can be described as follows:

$$\begin{aligned} P_t^+ &= \{p \in P | d(p, C_t) \geq 0\}, \\ P_t^- &= \{p \in P | d(p, C_t) < 0\}. \end{aligned}$$

For two discrete points on the ROC curve t_1 and t_2 , where $FPr_{t_1} < FPr_{t_2}$ (and consequently $FNr_{t_1} > FNr_{t_2}$), the relation between the corresponding Voronoi tes-

sellation is: $V(P_{t_1}^+) \subset V(P_{t_2}^+)$, with $\subset$ denoting a proper subset. As for any t we have $P = P_t^+ \cup P_t^-$ any change in P_t^+ immediately leads to a change in P_t^- , resulting in: $V(P_{t_2}^-) \subset V(P_{t_1}^-)$. We use the convention that with increasing value x of t for the operating point we are accepting more False Positives and with increasing value of y of t we are accepting more False Negatives.

From the above it follows that there is a set of points T containing all the discrete locations on the given classifier's ROC curve that correspond to the classifier's outcomes for different trade-offs. For each element in T the classifier's output is determined and therefore we know which elements are changing their class membership. If we increase the rate of False Positives then:

$$P_t^\Delta = P_{t+\epsilon}^+ - P_t^+.$$

In terms of the Voronoi visualization if we have two subsequent elements $t_1, t_2 \in T$ we have:

$$V(P_{t_2}^+) = V(P_{t_1}^+) \cup V(P_{t_2}^\Delta),$$

for some arbitrary small ϵ . By moving the operating point to higher values of FP the corresponding Voronoi cells are added to the region already displayed.

When decisions change at those discrete points t_1 and t_2 , then most likely only one or at most several data instances will be assigned a different label. The consequence of this operation is the change in the appearance of the visualized data elements (from TP to TN for example). Since the color represents the original labels and the shape the predicted label, the color of the elements changing the assigned label between steps t_1 and t_2 will not change, but the shape will. This change in appearance is easily noticeable if there are few data elements plotted. However, when the scatterplot becomes cluttered, the change is less apparent. In such cases the change in color of the Voronoi regions is obviously easier to notice for the user than just a change in label expressed by color or shape of the data elements. That, again, motivates the use of the colored background on the scatterplot for one of the classes.

In order to enable the expert to steer the classification model according to the desirable trade-off, we can interactively move the operating point along the ROC curve. We connect the interactive ROC curve to the visualizations of the scatterplots, as used in (Migut and Worring 2012). This is an instantiation of the **connect** interaction technique, as proposed by (Yi et al. 2007). Since we want to visually observe what effect the change of trade-off has on the classifier, we instantly visualize the Voronoi-based decision boundary for the adjusted operating point in all the scatterplots displayed.

Moreover, we integrate some additional interaction techniques, from the generic set in (Yi et al. 2007). The user is able to interactively change the dimensions on the scatterplot (**reconfigure**), so that he can examine all possible combinations of dimensions. We enable the user to highlight the element of interest in the scatterplot (**select**), resulting in a color change of the selected element. The user can also de-select an element if he is no longer interested in it. These techniques, together with the connect interaction technique, allow to see the relation between the decision boundary and the selected elements for all the visualized dimensions.

Table 1 The Voronoi based visualization of the decision boundary is supported by additionally adding color (and shape for points) to visualizations

	Scatterplot	Histogram	ROC
True positive	Red circle on white background	Red bin, positive height on white background	
True negative	Blue circle on white background	Blue bin, positive height on green background	
False positive	Red triangle on green background	Red bin, negative height on green background	White area under the curve
False negative	Blue triangle on white background	Blue bin, negative height on white background	Green area under the curve

Summary of the color and shape choices for the visualization are given in this table. In all visual components the consistent visualization choices are made for the four groups TP, TN, FP, and FN. The red color is associated with a positive class, and blue with a negative class

4.5 Visual correspondence between different visual components

To make the visual components correspond to each other, we use the same color for the corresponding concept visualized in the histograms, as we use in the scatterplot. Different from the original (Poulet 2008) histogram, as shown in Fig. 3a, in our implementation the original labels are represented in the color of the histograms' bins. The class membership, as assigned by the classifiers, is represented in the color of the background of the histograms.

To summarize: the color of the data points corresponds to the color of the bins and indicates the original label. The color of the background on the scatterplot corresponds with the background color of the histogram and indicates class membership as assigned by the classifier. Moreover, the misclassified data elements are assigned a different shape in the scatterplot (a triangle as opposed to circles for correctly classified elements). In the histogram the misclassifications are expressed by negative height for the bins. The visualization choices for the four groups True Positives, True Negatives, False Positives, and False Negatives are summarized in Table 1. Figure 4 shows that the Voronoi-based representation and the histogram of distances are complementary.

5 Visualization experiments

In the previous section we proposed an interactive visualization framework to explore the most interesting characteristics of the decision boundary. To illustrate how the framework can be applied, we conduct several visual experiments. Due to the subjective nature of the problem, we limit ourselves to the question how the decision boundary visualization together with the histogram of distances and the interactive ROC curve allow the user to perform the tasks listed in Sect. 2. The tool to perform the experiments is implemented in Protovis (Bostock and Heer 2009). Protovis is a free and open-source, Javascript and SVG based toolkit for web-native visualizations. The Voronoi tessellation implementation in Javascript is the (Fortune 1987)

Fig. 4 Screenshot of combined visualizations to explore and analyze multi-dimensional decision boundaries in 2D for the liver dataset. The components here are: the ROC curve, the histogram of the distances to the multi-dimensional decision boundary and a scatterplot showing labels and Voronoi-based representation of the multi-dimensional boundary. On the scatterplot: the original class membership is visualized in color; *red* objects are instances diagnosed with the liver disorder and *blue* objects are healthy instances; the *circular shape* indicates that the original label corresponds to the predicted label; the *triangular shape* indicates that the original label differs from the predicted label; the decision boundary visualized using Voronoi-based approximation shows the class membership as assigned by the multi-dimensional classifier; the regions belonging to one of the classes are filled with a *green color*; the regions not filled belong to the second class (Color figure online)

algorithm implemented by Raymond Hill.¹ The classifiers are trained in Matlab, using the Toolbox for Pattern Recognition PRTTools² and Data Description toolbox: ddtools.³ All the visualization elements are part of the user interface, as shown in Fig. 4.

5.1 Experimental setup

In this section we show the results of applying the proposed methodology to five different datasets. Each dataset is used to demonstrate different characteristics and applications of the proposed methodology.

¹ (<http://www.raymondhill.net/voronoi/rhill-voronoi.php>)

² (<http://prtools.org/>)

³ (<http://homepage.tudelft.nl/n9d04/ddtools.html>)

Table 2 Performance of the selected classifiers for the Diabetes and Liver dataset obtained with 10-fold cross-validation, with 3 repeats

Classifier	Cross-val error % ($\pm$ std) Diabetes	Liver
5 Nearest neighbors	0.28 ($\pm$ 0.01)	0.33 ($\pm$ 0.01)
Fisher	0.11 ($\pm$ 0.003)	0.22 ($\pm$ 0.001)
Support vector machine	0.23 ($\pm$ 0.005)	0.31 ($\pm$ 0.01)

For each of the datasets the following sequence of actions is performed. The dataset is divided into a training set (2/3) and a test set (1/3). Two dimensions for both training and test set, are first visualized using scatterplots. The axes of the scatterplots can be changed interactively, therefore the user can browse through the scatterplots to explore all the combinations of dimensions. Subsequently, the classifier is chosen, trained on the training set and applied to the test set. The Voronoi based decision boundary is visualized in the scatterplot of the currently chosen dimensions. Finally, the performance on the test set is visualized using an ROC curve, together with the current operating point.

We compare three classifiers of different types namely: five-nearest neighbors, Fisher and Support Vector Machine. The optimized 10-fold cross-validation error rates based on 3 repeats for all examined classifiers and for both datasets are listed in Table 2.

The classifier can be examined using the visualization of the Voronoi-based approximation of the decision boundary and through the manipulation of the operating point on the ROC curve. The expert using the system can tune the classifier to accept/reject a certain amount of positive/negative examples or can tune the classifier's threshold, according to the application's needs.

5.1.1 Datasets

The following datasets have been chosen to demonstrate the strengths and weaknesses of our methodology:

1. *Iris dataset*: a comprehensible dataset to illustrate all the components of the visualization technique. This is perhaps the best known database in the pattern recognition literature. The data set contains 3 classes of 50 instances, each described by 4 features, where each class refers to a specific type of iris flower. One class is linearly separable from the other two. The latter are not linearly separable from each other. Since we focus on two class problems, and we have chosen a dataset with three classes, we illustrate a case of classification of one of the classes that is not linearly separable from the other two.
2. *Liver-disorder* and *Diabetes datasets*: more complex datasets that are examples of expert's decision making problems. These datasets have a limited number of dimensions, but exhibit a complex relation between features and class membership. The Liver-disorders dataset from the UCI Machine Learning Repository⁴ consists of 345 objects described by 6 features. The objects are divided into two classes based

⁴ (<http://mllearn.ics.uci.edu/databases/liver-disorders/>)

on whether they do or do not have a liver disorder. The second dataset Diabetes comes from the UCI Machine Learning Repository⁵ and consists of 768 objects described by 8 features. The Diabetes dataset was used to forecast the onset of diabetes. The data is divided into classes based on whether the subject was tested positive or negative for diabetes. The diabetes dataset was also used in the study of (Poulet 2008).

3. *Breast cancer dataset*: high-dimensional dataset, where a dimension selection technique has to be applied.⁶ This dataset is 30-dimensional, which is too high for making browsing through all of 2D combinations ($(30^2 - 30)/2$ scatterplots) a feasible task. Browsing through all combinations of dimensions might also not be relevant, since some of the dimensions do not provide useful information for the diagnosis of the breast cancer. Therefore, we apply feature selection to select those dimensions that contribute the most to the class separation process as done by the classification procedure. We use the simple but effective, forward feature selection, which can be applied to the original feature space in order to perform classification in the reduced space more accurately than in the original space. Three features have been selected by the Forward feature selection procedure. The performance of the Fisher classifier using these three features is as good as using the original 30 features.
4. *Census Income dataset*: Dataset with a high amount of instances, where our technique fails, if no data transformation techniques are applied.⁷ The dataset is used to predict whether a person income exceeds \$ 50K/yr based on census data of 48,842 subjects described by 14 dimensions.

5.2 Results

In this section we show how the obtained visualizations and the functionality of the proposed methodology allow the user to perform the tasks stated in Sect. 2. First of all, guidelines are given how to read the visualizations, to prevent the misinterpretation of the proposed visual representation of the decision boundary.

Since the data is projected into two dimensions the multi-dimensional structure of the dataset is not preserved. For some combinations of dimensions the visualized representation of the decision boundary might be highly fragmented. In some cases, even though the boundary separates the data perfectly in the multidimensional space, the boundary might even be highly fragmented for all 2D projections of the features. That may wrongly be interpreted as overfitting of a classifier. This can not be avoided if we want to plot the boundary in only 2 dimensions and in relation to the original features. An expert should be aware of this and keep it in mind while exploring the dataset and the classifier using those visual representations.

⁵ (<http://mllearn.ics.uci.edu/databases/pima-indians-diabetes/>)

⁶ (<http://archive.ics.uci.edu/ml/datasets/Breast+Cancer>)

⁷ (<http://archive.ics.uci.edu/ml/datasets/Adult>)

Fig. 5 Several combination of dimensions for the Iris dataset **a** for the whole dataset, **b** for the zoomed-in area where the decision boundary is ‘de-fragmented’ with the results of the Fisher classifier with two classes (setosa versus other)

5.2.1 Analyze important dimensions (T1)

The visualization of pairs of dimensions allows to instantly identify which combinations of dimensions play an important role in the classification process. We illustrate how this task is performed using only separation and direction characteristics of our representation of the decision boundary. If the decision boundary is fairly simple, meaning that the amount of piece-wise linear elements constituting the decision boundary is limited, it implies that for this particular combination of dimensions the classifier separates data well in the multi-dimensional space.

Figure 5a shows an example of how the classifier separates the data. The boundary in 2D consists of one region for one class and a few ‘islands’ which belong to the same class, but in the 2D projection are located in the region of another class. As indicated before, this does not mean the classifier overfits. It means that elements included in the green region belong to one class according to the multi-dimensional classifier and elements located in the white region belong to the other class. Figure 5b shows the same scatterplots, but now only part of them. We zoomed-in onto the area of the plot in the neighborhood of the decision boundary. That allows the user to see what the fragmentation of the boundary exactly looks like. Zooming in helps when

Fig. 6 Several combinations of dimensions for the Diabetes dataset and the Fisher classifier. These dimensions are not easy to interpret, but we could zoom into the area of interest and still examine the relation between the data instances and the Fisher multi-dimensional decision boundary

Fig. 7 (i) and (ii) The combination of the dimensions selected by the Feature Forward selection for the Breast Cancer dataset. (i) These two dimensions are not representing the separation very well, while (ii) these two dimensions do. (iii) The Census Income dataset consist of two many instance to be represented on the scatterplot

the decision boundary is fragmented and you can see which elements belong to which class.

Figure 6 shows the more complex dataset, where it is more challenging to see how the data is separated. However, we can see that the combination of dimensions shown in (ii) indicates that at least one of the dimensions (f2) does not separate the data well. The same dimension in combination with f5, shown in (i), separates the data slightly better. The best separation is visible in (iii) and this scatterplot could be examined further, on the level of individual data elements.

Figure 7(i) and (ii) show that after dimension selection the same task can be performed for the high-dimensional dataset. The (i) scatterplot shows that these two dimensions do not separate data well, while dimensions presented in (ii) are worth further investigation.

Figure 7(iii) shows that for the dataset with too many instances, the cluttered scatterplot can not be analyzed. Not only the importance of the dimensions can not be assessed, but also none of the other tasks can be performed. The scatterplot is simply not readable.

5.2.2 Analyze and compare how different classifiers separate the data (T2)

Once we obtain a general idea about the importance of dimensions, we can compare those interesting dimensions for different classifiers. The visualization of the decision boundary in relation to the data makes it clear which data elements are classified correctly and which wrongly. Once we can visually examine which data objects are on which side of the decision boundary, we can easily see for which data objects the classifiers differ in assigning the label. The user can directly observe the behavior of a classifier. Moreover, the classifiers can be compared, allowing the user to inspect specific generalization characteristics. In fact, any inconsistently classified data elements by any of the compared classifiers could be instantly detected and analyzed in more detail. Therefore, the similarity of the models generated by different classifiers can be compared, providing more insight than just accuracy. That means that even though two classifiers might have similar performance in terms of accuracy, it might be favorable to choose one of the classifiers above the other, depending on specific needs/knowledge of the expert. The comparison can also be carried out based on the distance criterion. Since the elements closest to the boundary (distance) are the most probable to be classified differently by the different classifiers, we can closely observe their labels (separation and direction) on the visualizations. Using distance characteristics a group of suspect elements is highlighted, and using separation and direction characteristics the difficult ones (that are classified differently by different classifiers) can easily be found.

Figure 8 shows the decision boundary for two dimensions for different classifiers for the Liver dataset. It can be directly seen, that some data elements, are classified differently by the different classifiers. Some easily noticeable differences are in-lined. The elements in the ‘box’ are both correctly classified as a positive class by the Fisher and SVM classifiers, while the 5-nn classifier puts them both incorrectly in the negative class. Those elements are worth further investigation.

Figure 9 shows similar phenomena. Some elements are assigned a different label by the three classifiers presented. This could be valuable information when choosing a classifier for a specific application.

Fig. 8 Combination of two dimensions for the Liver dataset with the decision boundary produced by (i) Fisher, (ii) SVM, (iii) 5-nn classifier, showing that they separate the data in a fairly different way. For example, the two elements in the rectangle box are classified to the same class by Fisher and SVM classifiers, but to a different class by the 5-nn classifier

Fig. 9 Combination of two dimensions for the Diabetes dataset with the decision boundary produced by (i) Fisher, (ii) SVM, (iii) 5-nn classifiers, showing that they separate the data in a fairly different way

5.2.3 Analyze the relation between boundary and data elements (T3)

In order to analyze the data elements in relation to the decision boundary several interaction techniques are provided. First, the data element of interest can be highlighted and therefore can be traced in every plots, revealing all its characteristics. Its position with respect to the decision boundary can be established through the distance histogram. The interactively linked visualizations of the combinations of dimensions can be used to compare which labels are assigned to the same data elements by different classifiers. Therefore, we can observe which data points are difficult to model correctly. Those are the data points which, regardless of the performance of the classifiers, are being assigned a wrong label.

Figure 10 shows how this task can be performed. We took two scatterplots from Fig. 8, namely (ii)(a) and (ii)(b) and we zoomed in. On the histograms for both classifiers, we selected the correctly classified positive examples closest to the multi-dimensional decision boundary. Those elements are highlighted on the scatterplots. Therefore, we can compare on a high level of detail how elements are classified and how far they are from the decision boundary.

5.2.4 Analyze and compare classification costs (T4)

The costs for the current operating point of the classifier can be directly assessed through the classification error and observed on the visualizations of the decision boundaries. If we are not interested in the equal error rate, we might want to lower the number of false positives or on the contrary lower the number of false negatives. Since the operating point on the ROC curve is interactive, the costs of the classification can be instantly updated. This results in the immediate update in the decision boundary visualization and in the distance histograms. Figure 11 shows how this task can be performed. Once the operating point is changed, the visualization of the boundary changes. To explore the elements that are assigned a different label after changing the operating point, we can look into their details. The interactive operating point on the ROC curve is not a way of tuning the classifier, but examining how the classi-

Fig. 10 The LIVER dataset for dimensions f_3 and f_2 and (a) SVM and (b) Fisher classifiers. We zoom into the plots to explore a selected area of the data. On the histogram of the distances we highlight the bin corresponding to the correctly classified positive examples that are the closest to the multi-dimensional decision boundary. The elements corresponding to those distances are highlighted on the scatterplots

fier separates the data for each of the operating points. Therefore, depending to the application, the suitable balance between false positives and false negatives can be chosen.

5.3 Discussion

Through the visualization experiments we have shown how our approach to decision boundary visualization can be used for exploring the characteristics of multi-

Fig. 11 The Voronoi-based approximation of the decision boundary for the three different operating points for the Liver dataset and the SVM classifier and corresponding error rate: **a** minimizing FP rate, **b** current operating point chosen by the classifier, **c** maximizing TP rate

dimensional classifiers. We have also shown that for datasets with certain characteristics our technique is not sufficient on its own and extra measures have to be taken. In this section we summarize the current weaknesses of our method and address the possible solutions to overcome them. First, for a data set with a high number of dimensions it is not feasible to visualize and explore all the 2D combinations of dimensions. In those cases several techniques can be used as described in Sect. 2.2. We have shown that by using a simple feature selection algorithm the high-dimensional breast cancer dataset can also be analyzed well.

Second, even when the number of dimensions is limited, it is still likely that the user will need to search for the appropriate two dimensions. This could be solved by sorting the scatterplots according to the contribution of the dimensions in the classification process. In classifications terms, the weights of dimensions, as assigned by the classifier could be the indication of their importance for the user.

Third, given a large number of data samples, the scatterplot can become cluttered. This is an inherent problem of this visualization technique. A possible solution to this problem is to transform the data to a smaller representation. Several objects (prototypes) can be selected to best represent the dataset. Classification with prototype-based approach has been successfully adopted in machine learning and visualization (Pekalska et al. 2006), (Migut et al. 2011).

Finally, in our approach we limited ourself to two class problems. However, the proposed methodology could be translated to multi-class problems. The challenging part of this translation would be to represent the performance using an ROC curve for a multi-class classifier. For c classes this could be realized by using a series of c ROC curves, each showing one versus all other classes for one of the classes. In this way, the proposed methodology generalizes not only over the classifiers used, but also becomes dataset independent.

6 Conclusions

This paper proposes a method to visually represent a multi-dimensional decision boundary in 2D. We formalized the characteristics of the classifier, that should be captured by the visual representation of the multi-dimensional decision boundary in 2D, namely **Separation**, **Direction**, **Distance**, and **Performance**. We defined four tasks that can be performed by the expert: (1) analyze important dimensions, (2) compare different classifiers, (3) analyze the relation between the boundary and the data, and (4) compare classification costs. We thoroughly described why it is challenging to visually represent a multi-dimensional decision boundary in 2D, while complying with the classifier's characteristics and allowing execution of the defined tasks. To realize our idea, we developed a system that couples the visualization of the dataset, a Voronoi-based visualization of the multi-dimensional decision boundary in 2D, the histogram of the distances to the multi-dimensional decision boundary, and a visualization of the classifier's performance. We have shown, that using the Voronoi decomposition on two dimensions of classified data, we can visualize an approximation of the multi-dimensional decision boundary, expressing the two characteristics of the boundary: **Separation** and **Direction**. This visualization is an approximation of the actual decision boundary and does not represent absolute distances between the data elements and the decision boundary. We compensate for this by visualizing the distances using a histogram, expressing the **Distance** characteristic of the classifier. This combination of techniques allows the analysis of the classifier's behavior and it allows the visual assessment of the quality of the model. It also allows to examine characteristics of the dataset with respect to the classification model used. The proposed method is generic and can be used for different kinds of classifiers, allowing visual comparison among them. Moreover, such a visualized decision boundary can be explored for different trade-offs of the classifier by means of an ROC curve with an interactive operating point to study **Performance**. Through visual examples, we have shown that using this methodology we can perform the four tasks corresponding with the challenges of expert's decision making process. All the visualizations and interactions presented, contribute to the ultimate goal of being able to get insight in the classification problem at hand and use the insight to choose optimal classifiers.

Acknowledgments This research is supported by the Expertise center for Forensic Psychiatry, The Netherlands.

References

- Aurenhammer F (1991) Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Comput Surv* 23(3):345–405
- Bendix F, Kosara R, Hauser H (2005) Parallel sets: visual analysis of categorical data. In: *Proceedings of the 2005 IEEE Symposium on Information Visualization*, pp 133–140
- Bostock M, Heer J (2009) Protovis: a graphical toolkit for visualization. *IEEE Trans Vis Comput Graph (Proc InfoVis)* 15(6):1121–1128
- Brown ET, Liu J, Brodley CE, Chang R (2012) Dis-function: learning distance functions interactively. In: *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, (VAST)*, pp 83–92

- Caragea D, Cook D, Honavar VG (2001) Gaining insights into support vector machine pattern classifiers using projection-based tour methods. In: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 251–256
- Choo J, Lee H, Kihm J, Park H (2010) iVisClassifier: an interactive visual analytics system for classification based on supervised dimension reduction. In: Proceedings of the IEEE Symposium on Visual Analytics, Science and Technology, pp 27–34
- Cleveland W, McGill ME (1988) Dynamic graphics for statistics. Statistics/Probability Series
- Duda RO, Hart PE, Stork DG (2000) Pattern classification. Wiley-Interscience, Berlin
- Elmqvist N, Dragicevic P, Fekete JD (2008) Rolling the dice: multidimensional visual exploration using scatterplot matrix navigation. *IEEE Trans Vis Comput Graph (Proc InfoVis 2008)* 14(6):1141–1148
- Endert A, Han C, Maiti D, House L, Leman S, North C (2011) Observation-level interaction with statistical models for visual analytics. In: Proceedings of the IEEE Conference on Visual Analytics, Science and Technology, pp 121–130
- Fortune S (1987) A sweepline algorithm for Voronoi diagrams. *Algorithmica* 2(1–4):153–174
- Hamel L (2006) Visualization of support vector machines with unsupervised learning. In: Proceedings of 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp 1–8
- Heimerl F, Koch S, Bosch H, Ertl T (2012) Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, pp 2839–2848
- Hoferlin B, Netzel R, Hoferlin M, Weiskopf D, Heidemann G (2012) Inter-active learning of ad-hoc classifiers for video visual analytics. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST), pp 23–32
- Jeong DH, Ziemkiewicz C, Fisher B, Ribarsky W, Chang R (2009) iPCA: an interactive system for PCA-based visual analytics. *Comput Graph Forum* 28(3):767–774
- Keim DA (2002) Information visualization and visual data mining. *IEEE Trans Vis Comput Graph* 8(1):1–8
- Keim DA, Mansmann F, Schneidewind J, Thomas J, Ziegler H (2008) Visual analytics: scope and challenges. Springer, Berlin
- Malik A, Maciejewski R, Elmqvist N, Jang Y, Ebert D, Huang W (2012) A correlative analysis process in a visual analytics environment. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST), pp 33–42
- McClish DK (1989) Analyzing a portion of the ROC curve. *Med Decis Mak* 9(3):190–195
- Migut M, Worring M (2012) Visual exploration of classification models for various data types in risk assessment. *Inf Vis J (IVS)* 11(3):237–251
- Migut M, van Gemert J, Worring M (2011) Interactive decision making using dissimilarity to visually represented prototypes. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology (IEEE VAST), pp 141–149
- Pekalska E, Duin RPW, Paclík P (2006) Prototype selection for dissimilarity-based classifiers. *Pattern Recognit* 39(2):189–208
- Poulet F (2008) Towards effective visual mining with cooperative approaches. Springer, Berlin
- Provost F, Fawcett T (1997) Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp 43–48
- Stolte C, Tang D, Hanrahan P (2002) Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Trans Vis Comput Graph* 8(1):52–65
- Swayne DF, Lang DT, Buja A, Cook D (2003) GGobi: evolving from xgobi into an extensible framework for interactive data visualization. *Comput Stat Data Anal* 43(4):423–444
- Thomas J, Cook K (2005) Illuminating the path: the research and development agenda for visual analytics. IEEE CS Press, Silver Spring
- Ward MO (1994) Xmdvtool: Integrating multiple methods for visualizing multivariate data. In: Proceedings of the Conference on Visualization '94, pp 326–333
- Yan Z, Xu C (2008) Using decision boundary to analyze classifiers. 3rd International Conference on Intelligent System and Knowledge, Engineering 1:302–307
- Yi J, Kang J, Skasko J, Jacko J (2007) Toward a deeper understanding of the role of interaction in information visualization. *IEEE Trans Vis Comput Graph* 13(6):1224–1231