Visualizing multi-dimensional decision boundaries in 2D

Article *in* Data Mining and Knowledge Discovery · January 2013

DOI: 10.1007/s10618-013-0342-x

CITATIONS
10

READS
2,424

3 authors, including:

Marcel Worring
University of Amsterdam
308 PUBLICATIONS **14,794** CITATIONS

SEE PROFILE

Cor J. Veenman
Leiden University
58 PUBLICATIONS **2,532** CITATIONS

SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

Visualizing Multi-Dimensional Decision Boundaries in 2D

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

Received: date / Accepted: date

Abstract In many applications well informed decisions have to be made based on analysis of multi-dimensional data. The decision making process can be supported by various automated classification models. 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. We define a set of characteristics of the multi-dimensional classification model that have to be visually represented. To present those characteristics for both linear and non-linear methods, we combine visualization of the Voronoi based representation of multi-dimensional decision boundaries in scatterplots with visualization of the distances to the multi-dimensional boundary of all the data elements. We use interactive decision point selection on the ROC curve to allow the decision maker to refine the threshold of the classification model and instantly observe the results. 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

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 deepening

M.A. Migut
Intelligent System Lab Amsterdam, University of Amsterdam, The Netherlands
E-mail: mmigut@gmail.com

M.Worring
Intelligent System Lab Amsterdam, University of Amsterdam, The Netherlands
E-mail: M.Worring@uva.nl

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

the insight into the domain and it's underlying processes, intuitive means to assess the characteristics the classifier are important.

To support multi-dimensional decision making, it is favorable to obtain a visual comprehension of the classifier. Many visualization techniques are used 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 on individual data elements. Additionally, we need to assess, which elements are easily classified, which are more complex to handle, and what mistakes are made. This is an intricate interplay between the characteristics of the data, the classification model, and its parameter settings. For selecting the suitable classifier we need to go beyond performance as the only measure. 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, assuring critical cases are handled appropriately, and finding the balance between those mistakes.

One of the most informative characteristics of the classification model and its relation to the data is the decision boundary. The decision boundary 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 others have a higher certainty of class membership. Being able to visualize the decision boundary would be a great aid in decision making.

Decision boundaries can easily be visualized for 2D and 3D datasets (Duda et al, 2000). Generalizing beyond 3D forms a challenge in terms of the visualization and its use by the domain expert. The challenge in visualizing the decision boundary is that the boundary is defined in a multi-dimensional space. Transforming such a multi-dimensional boundary to a representation in lower dimensions, that can be displayed and understood by the experts is difficult. 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, 2010). 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 to relate the visualization of the decision boundary to the data elements in terms that are meaningful for the domain experts, in terms of class membership. The method in (Migut and Worring, 2010) does not allow 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, 2010; Poulet, 2008) to find a generic approach to 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, 2010) 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. The subsequent section presents a review of several attempts to visualize multi-dimensional decision boundaries. We pin-point the shortcomings of the existing techniques and propose to use the beneficial features of those. Then, we formally state the problem, we define a set of tasks that require a visual representation of the decision boundary, and propose a set of classifier 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. From there, we show how to interactively integrate several visualization techniques that contribute to a solution satisfying our requirements. In the following section we demonstrate the approach using two biomedical datasets, illustrating that the proposed methodology is suitable for exploring multi-dimensional classifiers.

2 Related work

Several 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, which means that they are not generic.

In (Poulet, 2008) authors display the histograms of the distance to the boundary distribution of correctly classified examples and misclassified examples for SVM. 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. The proposed method could also be applied to other classifiers like decision trees or regression lines. This is an interesting approach to decision boundary visualization offering a good estimation of the quality of the boundary. 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. Figure 3(a) and (b) show

an example using this technique for a combination of two arbitrary dimensions of an arbitrary dataset (LIVER dataset). 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 illustrate decision boundaries on SOM maps. The decision boundaries are not visualized in the original data space (domain space). 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.

3 Visualization of multi-dimensional decision boundaries

3.1 Problem analysis

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

Assume a training dataset with k objects represented by feature vectors with numerical data in an n-dimensional metric space. In this paper we only consider the two class problems. In section 5 however, we indicate how to transfer the techniques to multi-class problems.

Also, the features are limited to those which have meaning to the expert, so p is small. A classifier is trained on the dataset, resulting in a decision boundary in the n-dimensional space. 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 such an n-dimensional decision boundary to support the expert's decision making process. To that end, let us first look at the tasks that the expert has to perform. We summarize them as follows:

[T1] Task 1: Analyze which of the p 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 p-dimensional decision boundary a challenging task. There are three 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.

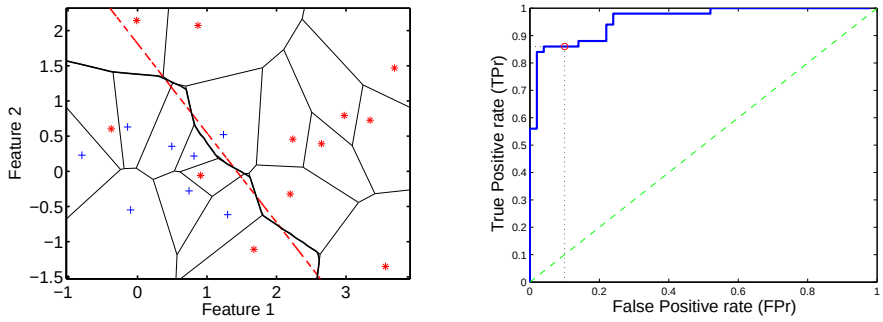
2. **Direction**: for two arbitrary data objects in the visualization, the representation of the decision boundary must unambiguously show on which side of the decision boundary each object is located in the multi-dimensional space.
3. **Distance**: for each visualized data object the distance to the decision boundary in multi-dimensional space should be represented.

To compare different classifiers, the visualization technique must be coherent and represent those three characteristics independently of the classification method used. Moreover, the data visualization technique must be chosen such that it allows the visualization of these characteristics and more importantly that it supports the conceptual framework of the experts. A taxonomy of 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). It is desirable to provide experts with an easy to understand visual representation of the data. We therefore choose the frequently used scatterplots. They are basic building blocks in statistical graphics and data visualization (Cleveland and McGill, 1988). Multidimensional visualization tools that feature scatterplots, such as Spotfire (Inc., 2007), 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. As 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), they accommodate the conceptual framework of the user. Moreover their familiarity among the 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 found in realistic datasets. Therefore, the scatterplots should be visualized for all combinations of the dimensions, where all the dimensions can be explored by the user. Different dimensions can be explored though, for example using an interactive axis, where the user selects the attribute to be plotted. In this way the variables are plotted against each other preserving the meaning of their values. Consequently, the decision boundary in multi-dimensional space has to be visualized in such a 2D setting.

3.2 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 classifier, however, can not be meaningfully projected into these two dimensions, as it is not defined in this 2D space. Hence, the projection into 2D will not represent the multi-dimensional classifier. The resulting projections do not necessary 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. Other linear boundaries, as well as non-linear boundaries, projected to 2D are meaningless. Straightforward projection of the classifier to 2D captures neither the **Separation** nor the **Direction**



(a) (b)

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.

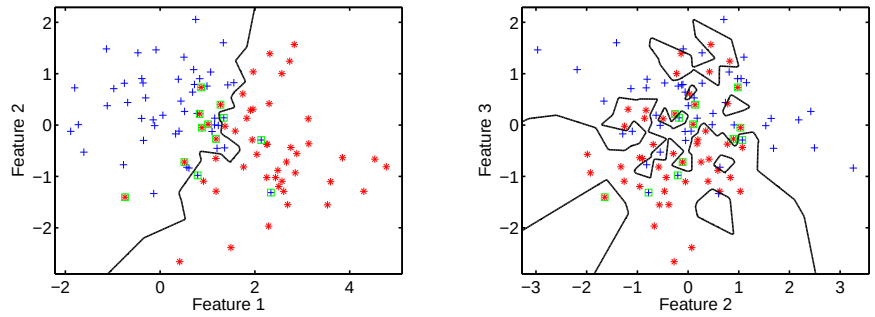
characteristics of the multi-dimensional classifier. The **Distance** characteristic is also not represented. Since the distances in 2D data projection do not represent the actual distances in the multi-dimensional space, the distances of data elements to the projected boundary would, therefore, also 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.

3.2.1 Voronoi based decision boundary visualization

In this section we describe the Voronoi-based representation of the decision boundary, as used in (Migut and Worring, 2010). 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).

Lets now consider two elements in the dataset, a labeled as positive by classifier B (boundary) and b labeled as negative by classifier B. The **Separation** characteristic of the classifier implies that in the visual representation of classifier B 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. Without knowledge of the actual distances it could be put midway the two elements. This resembles the Voronoi tessellation of the space, if performed for all the elements in the dataset. Therefore, we use the Voronoi tessellation to represent 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 figure 1. Formally, if P denotes a set of k-points, then



(a) (b)

Fig. 2 Decision boundary for a 10-dimensional dataset of two Gaussian distributed classes for the SVM classifier: (a)features well separated by the decision boundary (b)features highly fragmented by the decision boundary. Misclassified examples are marked with a green square.

for two distinct points $(p, q) \in P$ the separator separates all points of the plane closer to p from those closer to q:

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

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:

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

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. All the data objects are used as nodes to make the Voronoi diagram.

The boundaries of the Voronoi regions corresponding to neighbors belonging to different classes (according to the labels assigned by a 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 figure 1.

Visualization of the 2D combinations of the dimensions used by the classifier results in a series of scatterplots. Figure 2 illustrates our approach. 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.

To emphasize on which side of the boundary the elements are located, we color the Voronoi regions belonging to one class, as labeled by the multi-dimensional classifier. We argue, that for complex boundaries, a region based visualization makes it easier to comprehend to which class each data instance belongs. More motivation is given in the next section.

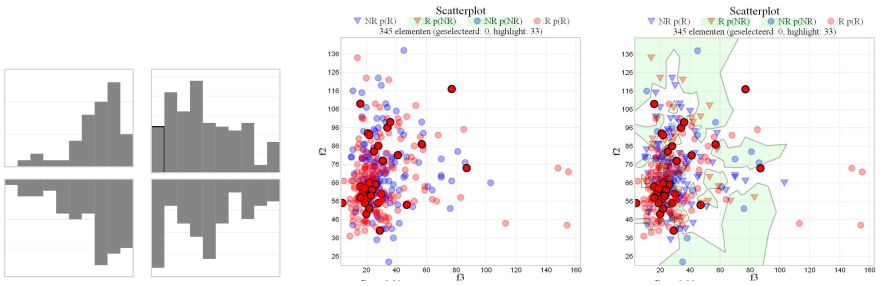
As stated in the previous subsection, the consequence of using scatterplots of multi-dimensional data is that the distances between data points from the original domain space are not preserved. Therefore, the distances between the data elements and the visualized decision boundary can also not be preserved. Using the Voronoi-based approach the distances between the actual objects and the decision boundary are indeed not preserved, but class membership is. 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 that 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 as proposed by (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 3(a) shows the four quarters. Histogram of distances allows making the difference in the distance to the multi-dimensional decision boundary visually distinctive adding to the "completeness" of decision boundary visualization. The class-membership and actual distances to the boundary can be explored. 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. The original labels are represented in color of the histograms' bins, while the classes, as assigned by the classifiers, are represented in the color of the background of the histograms. Figure 6 shows that the Voronoi-based representation and the histogram of distances are complementary.

3.3 Interactive trade-off inspection

In this section we show how to interactively couple trade-off visualizations with decision boundary visualizations, as proposed in Migut and Worring (2010).

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



(a) (b) (c)

Fig. 3 (a) The histogram of the distances of the TP, TN, FP, FN to decision boundary, with the highlighted bin of the closest TP to the boundary, as proposed in (Poulet, 2008) (b) The True Positive 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) The TP with the closest distance to the boundary are highlighted. Due to Voronoi based representation of decision boundary it is immediately visible how the boundary divides that 2D representation of the data.

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 example for this paper, we use the curve that represents the trade-off between the False Positives rate and False Negatives 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 B, let B_t be the decision boundary in n-dimensional space defined by the current operating point t. Further let $d(p, B_t)$ 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, a set of points classified to the positive class for the current operating point (P_t^+) can be described as follows:

$$P_t^+ = \{p \in P | d(p, B_t) \geq 0\}$$

$$P_t^- = \{p \in P | d(p, B_t) < 0\}$$

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 tessellation 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 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 Voronoi cells are added to the region displayed.

When decisions change at those discrete points t_1 and t_2, then most likely only one or at most several data instanced will be assigned a different label. 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.

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, 2010). 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, as proposed by (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.

4 Visualization experiments

In the previous section we propose 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 section 3.1. 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) algorithm im-

plemented by Raymond Hill [1]. The classifiers are trained in Matlab, using the Toolbox for Pattern Recognition PRTools[2] and Data Description toolbox: ddtools[3].

4.1 Experimental setup

For the visualization experiments we use two datasets, which are examples of expert's decision making problems. Those data sets (Liver-disorder and Diabetes) 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 Repository[4] consists of 345 objects described by 6 features. The objects are divided into two classes based on whether they do or do not have a liver disorder. The second dataset Diabetes comes from the UCI Machine Learning Repository[5] 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 object was tested positive or negative for diabetes. The diabetes dataset was also used in the study of (Poulet, 2008).

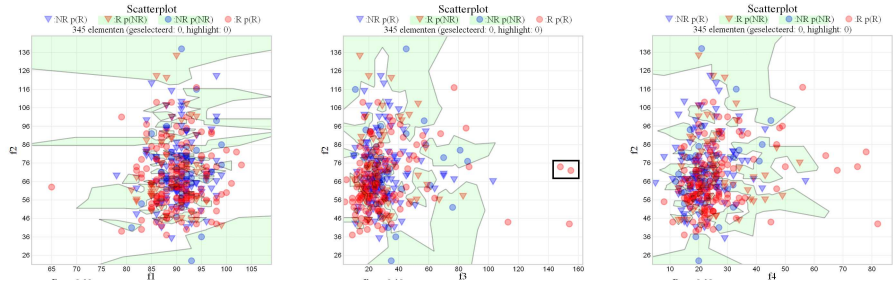
For each of the datasets the following sequence of actions is performed. The dataset is divided into a training set (2/3) and a test set (1/3). The two arbitrary dimensions for both training and test set, are first visualized using scatterplots. The axes of the scatterplots are interactive, therefore 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 currently chosen dimensions. The performance on the test set is visualized using ROC curve, together with the current operating point.

The classifiers we chose to compare are: 5-nearest neighbors, Fisher and Support Vector Machine (representing different types of classifiers). The optimized 10-fold cross-validation, over 3 repeats, error rates for all examined classifiers and for both datasets are listed in table 1. From a holistic point of view, the difference between the classifiers' performance is statistically insignificant. But they do not make the same mistakes. So it is worthwhile to explore which of the individual data instances are classified wrongly.

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. As an example, we present visualizations of the Voronoi-based approximation of the decision boundary for several combinations of the dimensions for the above mentioned classifiers for the LIVER dataset in figure 4 and for the DIABETES dataset in figure 5. All the dimensions can be explored, but we choose only a few combinations of dimensions to show what the visualizations look like.

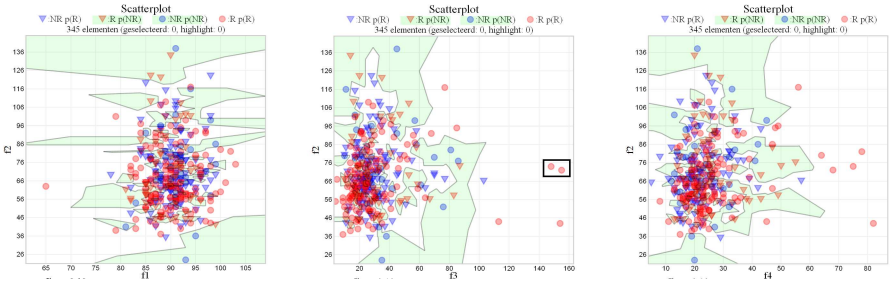
The operating point on the ROC curve can be manipulated for the classifier applied to the training set. Therefore, the expert using the system can tune the classifier to accept/reject a certain amount of positive/negative examples. He can tune the classifier's threshold, according to the application's needs.

[1] ('http://www.raymondhill.net/voronoi/rhill-voronoi.php')
[2] ('http://prtools.org/')
[3] ('http://homepage.tudelft.nl/n9d04/ddtools.html')
[4] ('http://mlearn.ics.uci.edu/databases/liver-disorders/')
[5] ('http://mlearn.ics.uci.edu/databases/pima-indians-diabetes/')



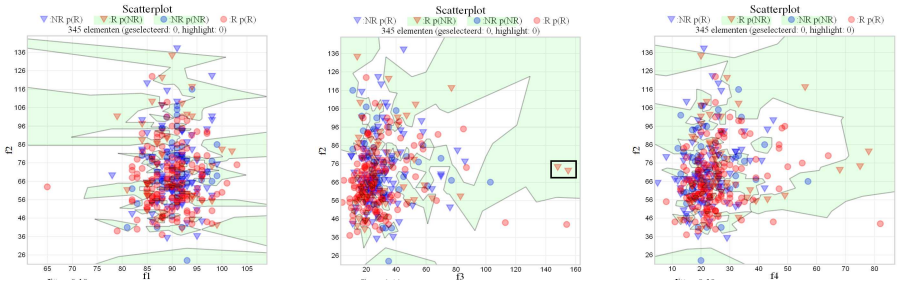
(i) (ii) (iii))

(a) SVM



(i) (ii) (iii)

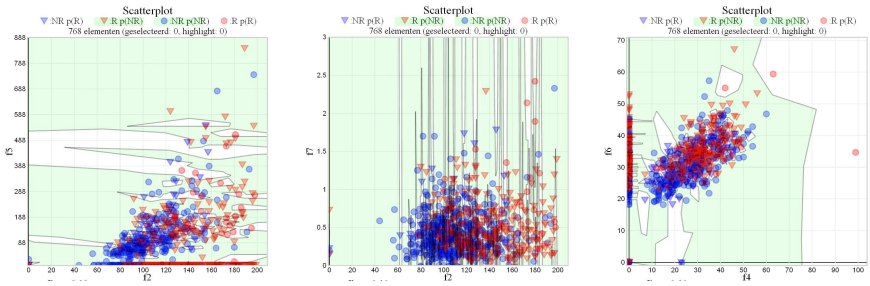
(b) FISHER



(i) (ii) (iii)

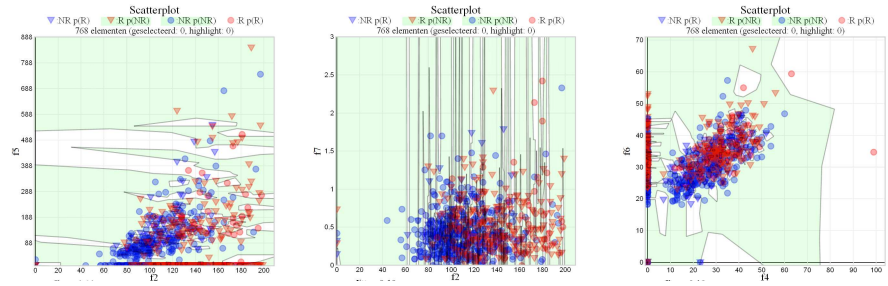
(c) 5-NN

Fig. 4 The decision boundary visualization for the LIVER dataset classified using (a) SVM, (b) Fisher classifier, (c) 5 nearest neighbor classifier. For each classifier the same set of dimensions have been chosen from all the possible combinations of dimensions that can be explored those are chosen for illustration purpose. 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.



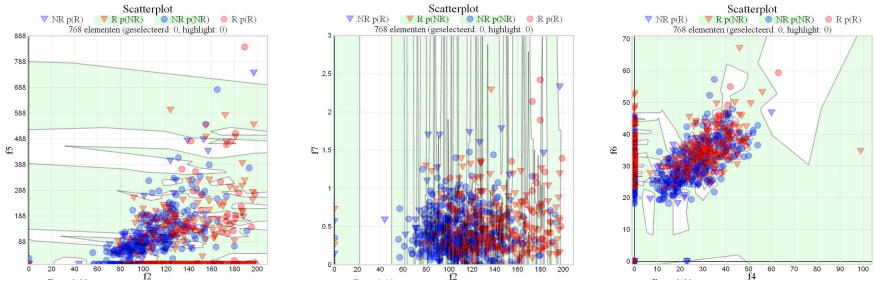
(i) (ii) (iii)

(a) SVM



(i) (ii) (iii)

(b) FISHER



(i) (ii) (iii)

(c) 5-NN

Fig. 5 The decision boundary visualization for the DIABETES dataset classified using (a) SVM, (b) Fisher classifier, (c) 5 nearest neighbor classifier. For more details on visualizations see the caption of figure 4.

Table 1 Performance of the selected classifiers for Diabetes and Liver dataset obtained with 10-fold cross-validation, with 3 repeats.

Classifier	Cross-val error% (±std) DIABETES	LIVER
5 Nearest Neighbors	0.28 (0.01)	0.33 (0.01)
Fisher	0.11 (0.003)	0.22 (0.001)
Support Vector Machine	0.23 (0.005)	0.31 (0.01)

4.2 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 section 3.1. First of all, guidelines are given how to read the visualizations, to prevent the misinterpretation of the proposed visual representation of the decision boundary.

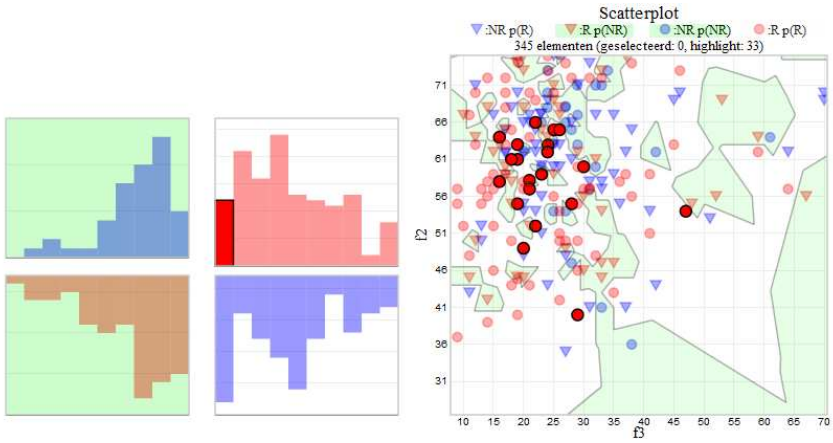
Since the data is projected into two dimensions the multi-dimensional structure of the dataset is not preserved. For some combinations of dimensions the visualized representation of the decision boundary might be highly fragmented. In some cases, even though the boundary separates the data perfectly 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.

4.2.1 Analyze important dimensions (T1)

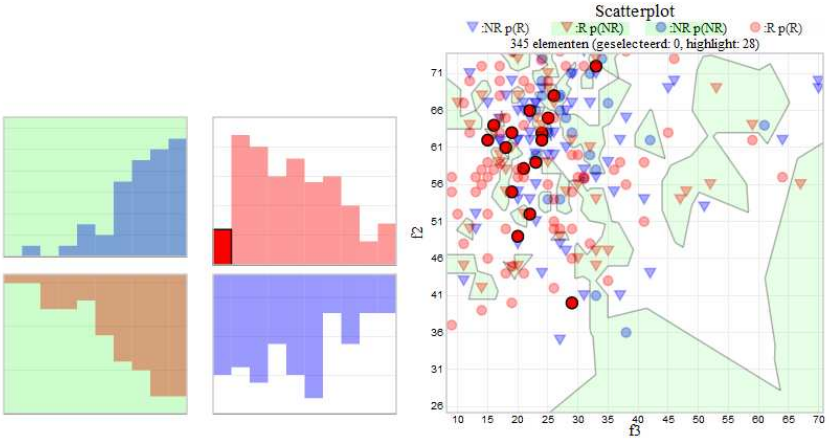
The visualization of pairs of dimensions allows to instantly identify which combination of dimensions play an important role in the classification process. If the decision boundary is fairly simple, meaning that the amount of piece-wise linear elements constituting the decision boundary is limited, it implies that for this particular combination of dimensions the classifier separates data well in the multi-dimensional space. Figure 5 illustrates how this task is performed using only the Voronoi based representation of the boundary. For each classifier we can directly conclude that the combination of dimensions shown in (ii) indicates that 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.

4.2.2 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. Any inconsistently classified data elements by any



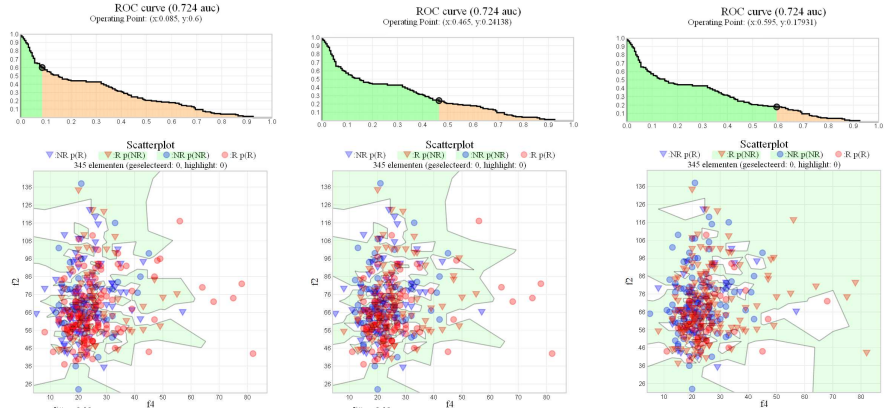
(a) SVM



(b) FISHER

Fig. 6 The LIVER dataset for dimensions f3 and f2 and (a) SVM and (b) Fisher classifiers. We zoom into the plots to explore the certain are 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.

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 an expert. Figure 4 illustrates how this task is performed. From the overview of the



(a) Error rate = 30% (b) Error rate = 37% (c) Error rate = 42%

Fig. 7 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.

combinations of dimensions for different classifiers, it can be directly seen, that some data elements, are classified differently by different classifiers. For example in (ii) for different classifiers, some easily noticeable differences are highlighted.

4.2.3 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 all 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 label is assigned to the same data elements by different classifiers. Therefore, we can observe which data points are difficult to learn correctly. Those are the data points which, regardless of the performance of the classifiers, are being assigned a wrong label. Figure 6 shows how this task can be performed. We took two scatterplots from figure 4, namely (ii)(a) and (ii)(b) and we zoomed in into these two plots. 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.

4.2.4 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

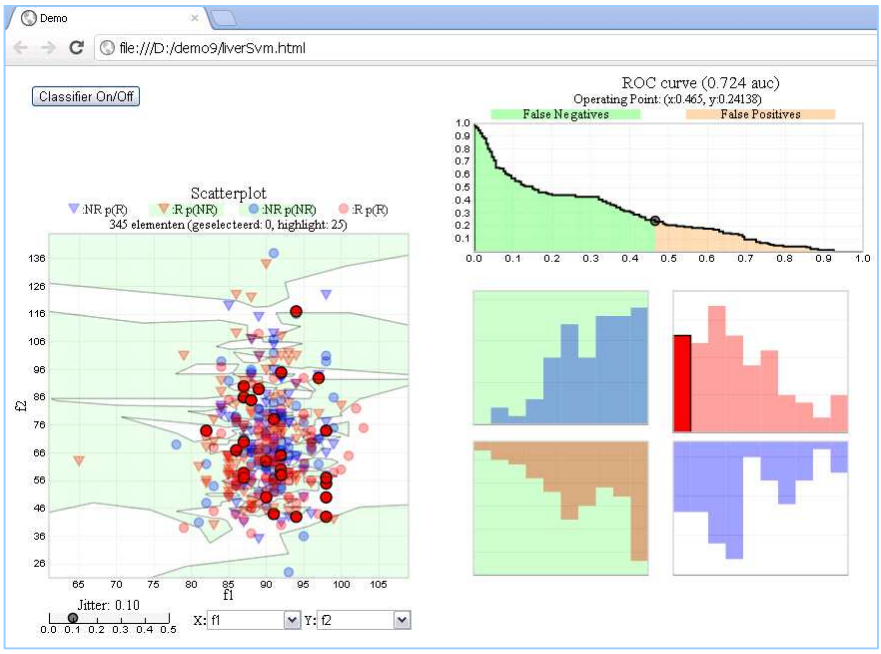


Fig. 8 Screenshot of combined visualizations to explore and analyze multi-dimensional decision boundaries in 2D. The components here are: the ROC curve, the histogram of the distances to the multi-dimensional decision boundary and a scatterplot showing labels and Voroni-based representation of multi-dimensional boundary.

can be instantly updated. This results in the immediate update in the decision boundary visualization and in the distance histograms. Figure 7 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 details of these points.

4.3 Framework

We have shown that to perform the defined tasks by exploring and analyzing the decision boundary of the classifier, we can use the Voronoi-based representation of the boundary combined with the interactive histogram of the distances to the multi-dimensional boundary and the ROC curve with an interactive operating point. Those elements should therefore be part of the user interface, e.g. as shown in figure 8.

5 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 decision boundary in 2D, namely **Separation**, **Direction**, and **Distance**. We defined four tasks that have to 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 decision boundary, 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 an 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. 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. In our approach we limited ourself to two class problems. However, the proposed methodology could be translated to the 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 for: one versus all other classes. In this way, the proposed methodology generalizes not only over the classifiers used, but also becomes dataset independent.

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.

Acknowledgements 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 COMPUTING SURVEYS 23(3):345–405

Bendix F, Kosara R, Hauser H (2005) Parallel sets: Visual analysis of categorical data. In: INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization

Bostock M, Heer J (2009) Protovis: A graphical toolkit for visualization. IEEE Trans Visualization & Comp Graphics (Proc InfoVis)

Caragea D, Cook D, Honavar VG (2001) Gaining insights into support vector machine pattern classifiers using projection-based tour methods. In: KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp 251–256

Cleveland W, McGill ME (1988) Dynamic graphics for statistics. Statistics/Probability Series

Duda RO, Hart PE, Stork DG (2000) Pattern Classification. Wiley-Interscience Publication

Fortune S (1987) A sweepline algorithm for voronoi diagrams. Algorithmica 2:153–174, URL http://dx.doi.org/10.1007/BF01840357, 10.1007/BF01840357

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

Inc S (2007) Spotfire. http://wwwspotfirecom

Keim DA (2002) Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics 8(1):1–8, DOI http://doi.ieeecomputersociety.org/10.1109/2945.981847

Keim DA, Mansmann F, Schneidewind J, Thomas J, Ziegler H (2008) Visual analytics: Scope and challenges pp 76–90

McClish DK (1989) Analyzing a portion of the ROC curve. Medical Decision Making 9(3):190–195

Migut M, Worring M (2010) Visual exploration of classification models for risk assessment. Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST) pp 11–18

Poulet F (2008) Towards Effective Visual Mining with Cooperative Approaches. Springer-Verlag, Berlin, Heidelberg

Provost F, Fawcett T (1997) Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: 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 Transactions on Visualization and Computer Graphics 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. Computational Statistics and Data Analysis 43(4):423–444

Thomas J, Cook K (2005) Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE CS Press

Ward MO (1994) Xmdvtool: integrating multiple methods for visualizing multivariate data. In: VIS '94: Proceedings of the conference on Visualization '94, IEEE Computer Society Press, 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, Stasko J, Jacko J (2007) Toward a deeper understanding of the role of interaction in information visualization. IEEE Transactions on Visualization and Computer Graphics 13(6):1224–1231