Charu C. Aggarwal

Data Classification: Algorithms and Applications

Contents

Chapter 1

Visual Classification

Giorgio Maria Di Nunzio
University of Padua
Italy
dinunzio@dei.unipd.it

1.1 Introduction

Extracting meaningful knowledge from very large datasets is a challenging task which requires the application of machine learning methods. This task is called data mining, the aim of which is to retrieve, explore, predict and derive new information from a given dataset. Given the complexity of the task and the size of the dataset, users should be involved in this process because, by providing adequate data and knowledge visualizations, the pattern

recognition capabilities of the human can be used to drive the learning algorithm [6]. This is the goal of Visual Data Mining: to present the data in some visual form, allowing the human to get insight into the data, draw conclusions, and directly interact with the data [18]. In [75], the authors define visual data mining as "the process of interaction and analytical reasoning with one or more visual representations of an abstract data that leads to the visual discovery or robust patterns in these data that form the information and knowledge utilised in informed decision making".

Visual data mining techniques have proven to be of high value in exploratory data analysis and they also have a high potential for exploring large databases [31]. This is particularly important in a context where an expert user could make use of domain knowledge to either confirm or correct a dubious classification result. An example of this interactive process is presented in [83], where the graphical interactive approaches to machine learning make the learning process explicit by visualizing the data and letting the user 'draw' the decision boundaries. In this work, parameters and model selection are no longer required because the user controls every step of the inductive process.

By means of visualisation techniques, researchers can focus and analyse patterns of data from datasets that are too complex to be handled by automated data analysis methods. The essential idea is to help researchers to examine the massive information stream at the right level of abstraction through appropriate visual representations and to take effective actions in real-time [47]. Interactive visual data mining has powerful implications in leveraging the intuitive abilities of the human for data mining problems. This may lead to solutions which can model data mining problems in a more intuitive and unrestricted way. Moreover, by using such techniques the user also has much better understanding of the output of the system even in the case of single test instances [1, 3].

The research field of Visual Data Mining has witnessed a constant growth and interest. In 1999, in a Guest Editor's Introduction of Computer Graphics and Application Journal [85], Wong writes:

> All signs indicate that the field of visual data mining will continue to grow at an even faster pace in the future. In universities and research labs, visual data mining will play a major role in physical and information sciences in the study of even larger and more complex scientific data sets. It will also play an active role in nontechnical disciplines to establish knowledge domains to search for answers and truths.

More than ten years later, Keim presents new challenges and applications [45]:

> Nearly all grand challenge problems of the 21st century, such as climate change, the energy crisis, the financial crisis, the health crisis and the security crisis, require the analysis of very large and complex datasets, which can be done neither by the computer nor the human alone. Visual analytics is a young active science field that comes with a mission of empowering people to find solutions for complex problems from large complex datasets. By tightly integrating human intelligence and intuition with the storage and processing power of computers, many recently developed visual analytics solutions successfully help people in analyzing large complex datasets in different application domains.

In this chapter, we focus on one particular task of visual data mining, namely visual classification. The classification of objects based on previously classified training data is an important area within data mining and has many real-world applications (see Section 1.3). The chapter is organized as follows: in this introduction, we present the requirements for Visual Classification (Section 1.1.1), a set of challenges (Section 1.1.3), and a brief overview of some of the approaches organized by visualisation metaphors (Section 1.1.2); in Section 1.2,

we present the main visualisation approaches for visual classification. For each approach, we introduce at least one of the seminal works and one application. In Section 1.3, we present some of the most recent visual classification systems which have been applied to real-world problems. In Section 1.4, we give our final remarks.

1.1.1 Requirements for Visual Classification

Shneidermann defines the "Visual Information Seeking Mantra" as a set of tasks that the user should perform [72]: overview first, zoom and filter, then details-on-demand. Along with this concept, the author proposes a type by task taxonomy of information visualizations. He lists seven tasks and seven data types. The tasks are: 1) to gain an overview of the entire collection; 2) zoom in on items of interest; 3) filter out uninteresting items; 4) select an item or group and get details when needed; 5) View relationships among items; 6) keep a history of actions to support undo, replay, and progressive refinement; 7) allow extraction of sub-collections and of the details when needed. The data types are: mono-dimensional, two-dimensional, three-dimensional, temporal, multi-dimensional, tree, network.

In [6], Ankerst and others discuss the reasons of involving the user in the process of classification: (i) by providing adequate data and knowledge visualizations, the pattern recognition capabilities of the human can be used to increase the effectivity of the classifier construction; (ii) the users have a deeper understanding of the resulting classifier; (iii) the user can provide domain knowledge to focus the learning algorithm better. Therefore, the main goal is to get a better cooperation between the user and the system: on the one hand, the user specifies the task, focuses the search, evaluates the results of the algorithm and feeds his domain knowledge directly into the learning algorithm; on the other hand, the machine learning algorithm presents patterns that satisfy the specified user constraints and creates appropriate visualizations.

In [9], a list of desired requirements for the visualization of the structure of classifiers are discussed. This list addresses specific requirements for what the users should be able to do when interacting with visual classification systems:

1. to quickly grasp the primary factors influencing the classification very little knowledge of statistics;
2. to see the whole model and understand how it applies to records, rather than the visualization being specific to every record;
3. to compare the relative evidence contributed by every value of every attribute;
4. to see a characterization of a given class, that is a list of attributes that differentiate that class from others;
5. to infer record counts and confidence in the shown probabilities so that the reliability of the classifier's prediction for specific values can be assessed quickly from the graphics;
6. to interact with the visualization to perform classification;
7. the system should handle many attributes without creating an incomprehensible visualization or a scene that is impractical to manipulate.

1.1.2 Visualization metaphors

Representing objects in two- or three-dimensional spaces is probably the most 'natural' metaphor a visualization system can offer to model object relationships. This is how we

perceive world as humans: two objects that are 'close' each other are probably more similar than two objects far away. The interactive visualization and navigation of such space becomes a means to browse and explore the dataset which match predetermined characteristics. In this section, we present a brief overview of some of the approaches covered in this chapter divided into two groups: approaches that represent objects using the metaphor of proximity to indicate similarity between objects in a two-dimensional or three-dimensional space, and other approaches which use more complex metaphors.

1.1.2.1 2D and 3D spaces

DocMINER [10] is a system which visualizes fine-granular relationships between single objects and allows the application of different object analysis methods. The actual mapping and visualization step uses a Self-Organizing Map (see Section 1.2.5.1). Given a distance metric, objects are mapped into a two-dimensional space, so that the relative error of the distances in this 2D space regarding the true distances of the objects is minimized. High-dimensional data sets contain several attributes, and finding interesting projections can be a difficult and time-consuming task for the analyst, since the number of possible projections increases exponentially with the number of concurrently visualized attributes. VizRank [53] is a method based on K-Nearest Neighbor distance [16] which is able to rank visual projections of classified data by their expected usefulness. Usefulness of a projection can then be defined as a property that describes how well clusters with different class values are geometrically separated. The system Bead [14] represents objects as particles in a three-dimensional space and the relationships between objects are represented by their relative spatial positions. In Galaxies [84], clusters of documents are displayed by reducing the high dimensional representation to a three-dimensional scatterplot. The key measurement for understanding this visualization is the notion of document similarity. ThemeScapes [84] is a three dimensional plot that mimics terrain topology. The surface of the terrain is intended to convey relevant information about topics and themes found within the corpus: elevation depicts theme strength, while valleys, cliffs and other features represent relationships between documents. In [69], authors present the use of three-dimensional surfaces for visualizing the clusters of the results of a search engine. The system lets users examine resulting three-dimensional shapes and immediately see differences and similarities in the results. Morpheus [62] is a tool for an interactive exploration of clusters of objects. It provides visualization techniques to present subspace clustering results such that users can gain both an overview of the detected patterns and the understanding of mutual relationships among clusters.

1.1.2.2 More complex metaphors

MineSet [9] provide several visualization tools that enable users to explore data and discover new patterns. Each analytical mining algorithm is coupled with a visualization tool that aids users in understanding the learned models. Perception-Based Classification [6] (PBC) was introduced as an interactive decision tree classifier based on a multidimensional visualization technique. The user can selected the split attributes and split points at each node and thus constructed the decision tree manually (see Section 1.2.6.1). This technique not only depicts the decision tree but it also provides explanations why the tree was constructed this way. The Evidence Visualizer [8] can display Bayes model decisions as pies and bar charts. In particular, the rows of pie charts represent each attribute, and each pie chart represents an interval or value of the attribute. ExplainD [67] is a framework for explaining decisions made by classifiers that use additive evidence. It has been applied to different linear model such as support vector machines, logistic regression and Naïve Bayes. The main goal of this framework is to visually explaining the decisions of machine-learned

classifiers and the evidence for those decisions. The Class Radial Visualization [70] is an integrated visualization system that provides interactive mechanisms for a deep analysis of classification results and procedures. In this system, class items are displayed as squares and equally distributed around the perimeter of a circle. Objects to be classified are displayed as colored points in the circle and the distance between the point and the squares represent the uncertainty of assigning that object to the class. In [66], authors presents two interactive methods to improve the results of a classification task: the first one is an interactive decision tree construction algorithm with a help mechanism based on Support Vector Machines (SVM); the second one is a visualization method used to try to explain SVM results. In particular, it uses a histogram of the data distribution according to the distance to the boundary and linked, a set of scatter-plot matrices or the parallel coordinates. This method can also be used to help the user in the parameter tuning step of SVM algorithm and reduce significantly the time needed for the classification.

1.1.3 Challenges in Visual Classification

In [45], Keim and others discuss the challenges of the future of visualization systems. Even though each individual application and task has its own requirements and specific problems to solve, there are some common challenges that may be connected to the task of visual classification. The challenges are six: scalability, uncertainty, hardware, interaction, evaluation, infrastructure. Scalability is probably one of the most important future challenges with the forthcoming 'era of big data'. Visual solution needs to scale in size, dimensionality, data types, and levels of quality. The relevant data patterns and relationships need to be visualized on different levels of details, and with appropriate levels of data and visual abstraction. Dealing with uncertainty in visual analytics is nontrivial because of the large amount of noise and missing values. The notion of data quality and the confidence of the algorithms for data analysis need to be appropriately represented. The analysts need to be aware of the uncertainty and be able to analyze quality properties at any stage of the data analysis process. Efficient computational methods and powerful hardware are needed to support near real time data processing and visualization for large data streams. In addition to high-resolution desktop displays, advanced display devices such as large-scale power walls and small portable personal assistants need to be supported. Visual analytics systems should adapt to the characteristics of the available output devices, supporting the visual analytics workflow on all levels of operation. Novel interaction techniques are needed to fully support the seamless intuitive visual communication with the system. User feedback should be taken as intelligently as possible, requiring as little user input as possible. A theoretically founded evaluation framework needs to be developed to assess the effectiveness, efficiency and user acceptance of new visual analytics techniques, methods, and models. For a deeper analysis of these challenges, we suggest [47].

Interaction, evaluation and infrastructure have been recently discussed in the ACM International Conference of Tabletops and Surfaces. In [54], the authors present the development of novel interaction techniques and interfaces for enhancing collocated multiuser collaboration so as to allow multiple users to explore large amounts of data. They build case studies where multiple users are going to interact with visualizations of a large data set like biology datasets, social networks datasets, and spatial data.

1.1.4 Related works

In this section, we want to give the reader some complementary readings about surveys on visual data mining. Compared to our, these surveys have different objectives and do not focus on the specific problem of visual classification. These surveys discuss issues that are

very important but go beyond the scope of this chapter. For example: how to choose the appropriate visualization tool, advantages and disadvantages, strength and weaknesses of each approach, how to extend basic visualization approaches.

In [18], an overview of the techniques available under the light of different categorizations is presented. The role of interaction techniques is discussed, as well as the important question of how to select an appropriate visualization technique for a task.

The problem of identifying adequate visual representation is also discussed in [57]. The authors classify the visual techniques in two classes: technical and interactive techniques. For each approach they discuss advantages and disadvantages in visualizing data to be mined.

[11] presents how to integrate visualization and data mining techniques for knowledge discovery. In particular, this work looks at strengths and weaknesses of information visualization techniques and data mining techniques.

In [25], the authors present a model for hierarchical aggregation in information visualization for the purpose of improving overview and scalability of large scale visualization. A set of standard visualization techniques is presented and a discussion of how they can be extended with hierarchical aggregation functionality is given.

1.2 Approaches

In this section, we present an overview of many of the most important approaches used in data visualization that have been applied to visual classification. This survey is specifically designed to present only visual classification approaches. For each approach, we added a reference to at least one of the seminal works and one example of an application for the specific classification task. We did not enter into discussions on the appropriateness, advantages and disadvantages of each technique, which can be found in other surveys presented in Section 1.1.4. We present the approaches in alphabetical order: nomograms, parallel coordinates, radial visualisations, scatter plots, topological maps, and trees. All the figures in this Section were produced with R[1], and the code to reproduce these plots can be freely downloaded.[2]

1.2.1 Nomograms

A nomogram [3] is any graphical representation of a numerical relationships. Invented by French mathematician Maurice d'Ocagne in 1891, the primary means of a nomogram was to enable the user to graphically compute the outcome of an equation without doing any calculus. Today, nomograms are often used in medicine to predict illness based on some evidence. For example, [58] shows the utility of such a tool to estimate the probability of diagnosis of acute myocardial infarction. In this case, the nomogram is designed in such a way that it can be printed on paper and easily used by physicians to obtain the probability of diagnosis without using any calculator or computer. There are a number of nomograms used in daily clinical practice for prognosis of outcomes of different treatments especially in

[1] http://www.r-project.org/
[2] http://www.purl.org/visualclassification
[3] http://archive.org/details/firstcourseinnom00broduoft

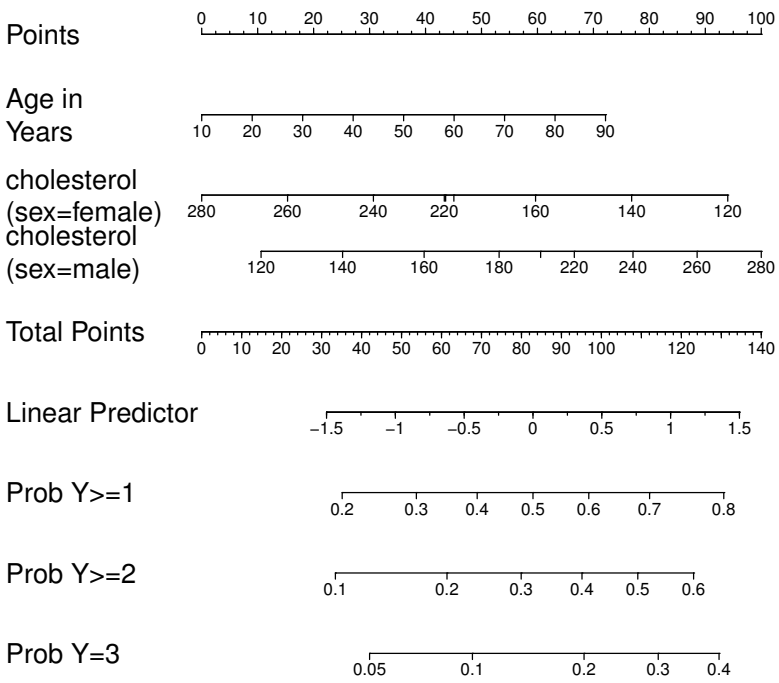

FIGURE 1.1: Nomograms. Given the age of a person and the level of cholesterol of the patient, by drawing a straight line that connects these two points on the graph, it is possible to read many information about survival probabilities ($Y >= 1$, $Y >= 2$, $Y = 3$) according to different combinations of features.

the field of oncology [44, 81]. In Figure 1.1, an example of a nomogram for predicting the probability of survival given factors like age and cholesterol level is shown. [4]

The main benefit of this approach is simple and clear visualization of the complete model and the quantitative information it contains. The visualization can be used for exploratory analysis and classification, as well as for comparing different probabilistic models.

1.2.1.1 Naïve Bayes Nomogram

In [61], the authors propose the use of nomograms to visualize Naïve Bayes classifiers. This particular visualisation method is appropriate for this type of classifiers since it clearly exposes the quantitative information on the effect of attribute values to class probabilities by using simple graphical objects (points, rulers and lines). This method can also be used to reveal the structure of the Bayes classifier and the relative influences of the attribute values to the class probability and to support the prediction.

[4] `http://cran.r-project.org/web/packages/rms/index.html`

1.2.2 Parallel Coordinates

Parallel coordinates have been widely adopted for the visualization of high-dimensional and multivariate datasets [39, 38]. By using parallel axes for dimensions, the parallel coordinates technique can represent n-dimensional data in a 2-dimensional space; consequently, it can be seen as a mapping from the space R^n into the space R^2. The process to project a point of the n-dimensional space into the 2-dimensional space is the following: on a two-dimensional plane with cartesian coordinates, starting on the y-axis, n copies of the real line are placed parallel (and equidistant) to the y-axis. Each line is labeled from x_1 to x_n. A point c with coordinates $(c_1, c_2, ..., c_n)$ is represented by a polygonal line whose n vertices are at $(i-1, c_i)$ for $i = 1, ..., n$.

Since points that belong to the same class are usually close in the n-dimensional space, objects of the same class have similar polygonal lines. Therefore, one can immediately see groups of lines that correspond to points of the same class. Axes ordering, spacing and filtering can significantly increase the effectiveness of this visualization, but these processes are complex for high dimensional datasets [86]. In [78], the authors present an approach to measure the quality of the parallel coordinates view according to some ranking functions.

In Figure 1.2, an example of parallel coordinates to classify the Iris Dataset [5] is shown. The four-dimensional object has been projected onto four parallel coordinates. Flowers of the same kind show similar polygonal patterns; however, edge cluttering is already a problem even with this small number of objects. [6]

1.2.2.1 Edge Cluttering

Although parallel coordinates is a useful visualization tool, edge clutter prevents effective revealing of underlying patterns in large datasets [89]. The main cause of the visual clutter comes from too many polygonal lines. Clustering lines is one of the most frequently used methods to reduce the visual clutter and improve the perceptibility of the patterns in multivariate datasets. The overall visual clustering is achieved by geometrically bundling lines and forming patterns. The visualization can be enhanced by varying color and opacity according to the local line density of the curves.

Another approach to avoid edge cluttering is angular histogram [29]. This technique considers each line-axis intersection as a vector, then both the magnitude and direction of these vectors are visualised to demonstrate the main trends of the data. Users can dynamically interact with this new plot to investigate and explore additional patterns.

1.2.3 Radial Visualizations

A radial display is a visualization paradigm in which information is laid out on a circle, ellipse, or spiral on the screen. Perhaps, the earliest use of a radial display in statistical graphics was the pie chart. However, the pie chart has some limitations. In particular, when the wedges in a pie chart are almost the same size, it is difficult to determine visually which wedge is largest. A bar chart is generally better suited for this task. For example, the Evidence Visualizer [13, 8] can display Bayes model decisions as pies and bar charts. In particular, the rows of pie charts represent each attribute, and each pie chart represents an interval or value of the attribute.

Many radial techniques can be regarded as projections of a visualization from a Cartesian coordinate system into a polar coordinate system. The idea behind a radial visualization is similar to the one of parallel coordinates; however, while the space needed for Parallel Co-

[5] `http://archive.ics.uci.edu/ml/datasets/Iris`
[6] `http://cran.r-project.org/web/packages/MASS/index.html`

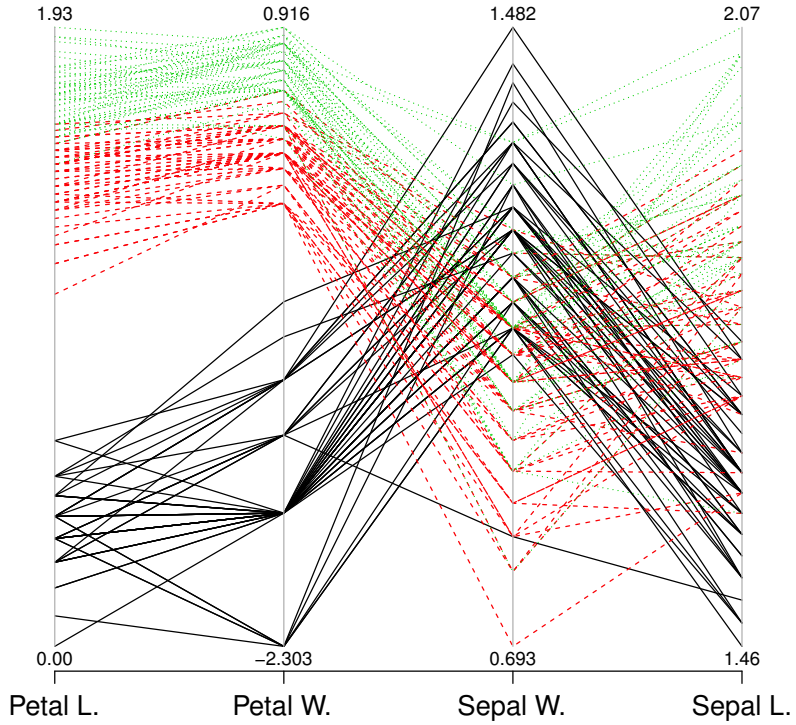

FIGURE 1.2: Parallel Coordinates. In this example, each object has four-dimensions and represent the characteristic of a species of Iris flower (petal and sepal width an length in logarithmic scale). The three types of lines represent the three kinds of Iris. With parallel coordinates, it is easy to see common patterns among flowers of the same kind; however, edge cluttering is already visible even with a small dataset.

ordinates increases with the number of dimensions, the space used by a radial visualisation remains fixed by the area of the circle. An example is Radviz [35, 36] where n-dimensional objects are represented by points inside a circle. The visualized attributes correspond to points equidistantly distributed along the circumference of the circle. [24] presents a survey on radial visualisation, while [22] discusses advantages and drawbacks of these methods compared to classical Cartesian visualisation.

1.2.3.1 Star Coordinates

Star Coordinates represents each dimension as an axis radiating from the center of a circle to the circumference of the circle [41]. A multi-dimensional object is mapped onto one point on each axis based on its value in the corresponding dimension. StarClass [79] is a visualisation tool allows users to visualize multi-dimensional data by projecting each data object to a point on 2D display space using Star Coordinates.

In Figure 1.3, three five-dimensional objects are mapped on a star coordinate plot. Each

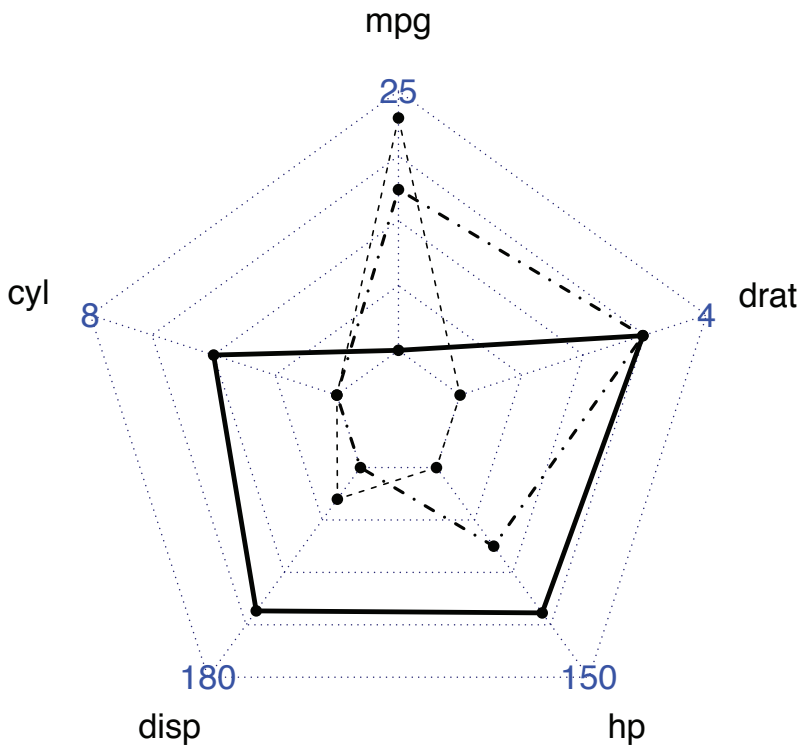

FIGURE 1.3: Star Coordinates. Three five-dimensional objects are mapped on this star plot. Each coordinate is one of the axis radiating from the center. In this examples, the three objects are cars described by features like: number of cylinders, horse power, miles per gallon. The number at the end of each axis represent the maximum value for that dimension. Cars with similar features, have similar polygons too.

coordinate is one of the axis radiating from the center. Objects that are similar in the original pace, have similar polygons too.

Star coordinates have been successfully used in revealing cluster structures. In [88, 87], an approach called Hypothesis Oriented Verification and Validation by Visualization (HOV) ioffers a tunable measure mechanism to project clustered subsets and non-clustered subsets from a multidimensional space to a 2D plane. By comparing the data distributions of the subsets, users not only have an intuitive visual evaluation but also have a precise evaluation on the consistency of cluster structure by calculating geometrical information of their data distributions.

1.2.4 Scatter Plots

Scatter plots use Cartesian coordinates to display the values of two- or three-dimensional data. Since most problems in data mining involve data with a large number of dimensions, dimensionality reduction is a necessary step to use this type of plots. Reduction can be performed by keeping only the most important dimensions, that is only those that hold

the most information and by projecting some dimensions onto others. The reduction of dimensionality can lead to an increased capability of extracting knowledge from the data by means of visualization, and to new possibilities in designing efficient and possibly more effective classification schemes [82]. A survey on the methods of dimension reduction that focus on visualizing multivariate data can be found in [26].

In Figure 1.4, a matrix of scatterplots shows all the possible combinations of features for the Iris Dataset. Even though the three species of flowers are not linearly separable, it is possible to study what pairs of features allow for a better separation. Even with this relatively few number of items, the problem of overlapping points is already visible.

In [46], the authors discuss the issue of the high degree of overlap in scatter plots in exploring large data sets. They propose a generalization of scatter plots where the analyst can control the degree of overlap allowing the analyst to generate many different views for revealing patterns and relations from the data. In [78], an alternative solution to this problem is given by presenting a way to measure the quality of the scatter plots view according to some ranking functions. For example, a projection into a two-dimensional space may need to satisfy a certain optimality criterion that attempts to preserve distances between the class-means. In [20], a kind of projections that are similar to Fishers linear discriminants, but faster to compute, are proposed. In [7], a type of plot which projects points on a two-dimensional plane called similarity-dissimilarity plot is discussed. This plot provides information about the quality of features in the feature space and classification accuracy can be predicted from the assessment of features on this plot. This approach has been studied on synthetic and real life datasets to prove the usefulness of the visualisation of high dimensional data in biomedical pattern classification.

1.2.4.1 Clustering

In [19], the authors compare two approaches for projecting multidimentional data onto a two-dimensional space: Principal Component Analysis (PCA) and random projection. They investigate which of these approaches fits best nearest neighbour classification when dealing with two types of high-dimensional data: images and micro arrays. The result of this investigation is that PCA is more effective for severe dimensionality reduction, while random projection is more suitable when keeping a high number of dimensions. By using one of the two approaches, the accuracy of the classifier is greatly improved. This shows that the use of PCA and random projection, may lead to more efficient and more effective, nearest neighbour classification. In [71], an interactive visualisation tool for high-speed power system frequency data streams is presented. A k-median approach for clustering is used to identify anomaly events in the data streams. The objective of this work is to visualize the deluge of expected data streams for global situational awareness, as well as the ability to detect disruptive events and classify them. [2] discusses a interactive approach for nearest neighbor search in order to choose projections of the data in which the patterns of the data containing the query point are well distinguished from the entire data set. The repeated feedback of the user over multiple iterations is used to determine a set of neighbors which are statistically significant and meaningful.

1.2.4.2 Naïve Bayes Classification

Naïve Bayes classifiers are one of the most used data mining approaches for classification. By using Bayes' rule, one can determine the posterior probability $Pr(c|x)$ that an object x belong to a category c in the following way:

$$Pr(c|x) = \frac{Pr(x|c)Pr(c)}{Pr(x)} \tag{1.1}$$

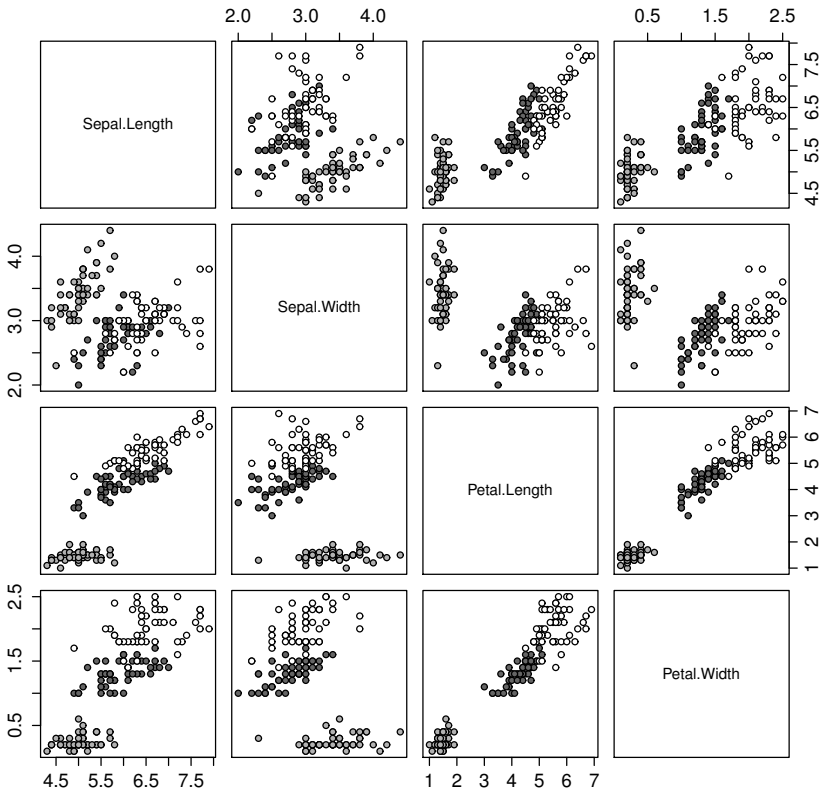

FIGURE 1.4: Scatter plots. In the Iris Dataset, flowers are represented by four-dimensional vectors. In this figure, the matrix of scatterplots presents all the possible two-dimensional combinations of features. The shade of grey of each point represents the kind of Iris. Some combinations allows for a better separation between classes; nevertheless, even with this relatively few number of items, the problem of overlapping points is already visible.

where $Pr(x|c)$ is the likelihood function, $P(c)$ the prior probability of the category c, and $P(x)$ the probability of an object x.

'Likelihood projections' is an approach which uses the likelihood function $P(x|c)$ for nonlinear projections [76]. The coordinates of this 'likelihood space' are the likelihood functions of the data for the various classes. In this new space, the Bayesian classifier between any two classes in the data space can be viewed as a simple linear discriminant of unit slope with respect to the axes representing the two classes. The key advantage of this space is that we are no longer restricted to considering only this linear discriminant. Classification can now be based on any suitable classifier that operates on the projected data. In [68], the likelihood space is used to classify speech audio. The projection of the audio data results in the transformation of diffuse, nebulous classes in high-dimensional space into compact clusters in the low-dimensional space that can be easily separated by simple clustering mechanisms. In this space, decision boundaries for optimal classification can be more easily identified using simple clustering criteria

In [21], a similar approach is used as a visualization tool to understand the relationships between categories of textual documents, and to help users to visually audit the classier and identify suspicious training data. When plotted on the Cartesian plane according to this formulation, the documents that belong to one category have specific shifts along the x-axis and the y-axis. This approach is very useful to compare the effect of different probabilistic models like Bernoulli, multinomial or Poisson. The same approach can be applied to the problem of parameters optimization for probabilistic text classifiers, as discussed by [63].

1.2.5 Topological Maps

Topological maps are a means to project an n-dimensional input data into a two-dimensional data by preserving some hidden structure or relation among data [49]. The automatic systems which make this projection can automatically form two- or three-dimensional maps of features that are present in sets of input signals. If these signals are related metrically in some structured way, the same structure will be reflected in the low dimensional space. In [64], the authors show how traditional distance-based approaches fail in high-dimensional spaces and propose a framework that supports topological analysis of high dimensional document point clouds. They describe two-stage method for topology-based projections from the original high dimensional information space to both 2D and 3D visualizations.

1.2.5.1 Self-Organizing Maps

A Self-Organizing Map (SOM) is a kind of neural network which preserve the topological properties of the input space by means of a neighborhood function [50]. It consists of units arranged as a two-dimensional or hexagonal grid where each unit represent a vector in the data space. During the training process, vectors from the dataset are presented to the map in random order and the unit with the highest response to a chosen vector and its neighborhood are adapted in such a way as to make them more responsive to similar inputs. SOMs are very useful for visualising multidimensional data and the relationships among the data on a two-dimensional space. For example, [60] presents a typical result from the application of self-organizing maps to the problem of text classification. The grid of units represent the document collection, each unit being a class. Once the network has been trained, the grid shows how the different classes are 'similar' to each other in terms of the distance on the grid.

In Figure 1.5, the result of the training of a SOM on a dataset of wines is shown. Each wine, described by a vector of thirteen features, has been projected on a 5 by 5 hexagonal grid. The shape of each point (triangle, circle, cross) represent the original category of the wine, the shade of grey of each activation unit is the predicted label. [7]

Recently, SOMs have been used to study weather analysis and prediction. Since weather patterns have a geographic extent, weather stations that are geographically close to each other should reflect these patterns. In [28], data collected by weather stations in Brazil are analysed to find weather patterns. Another important field of application of SOMs is DNA classification. In [59], the authors present an application of the hyperbolic SOM, a Self-Organizing Maps which visualises results on a hyperbolic surface. A hyperbolic SOM can perform visualisation, classification and clustering at the same time as a SOM; hyperbolic SOMs have the potential to achieve much better low-dimensional embeddings, since they offer more space due to the effect, that in a hyperbolic plane the area of a circle grows asymptotically exponential with its radius. Moreover, it also incorporates links between

[7] `http://cran.r-project.org/web/packages/som/`

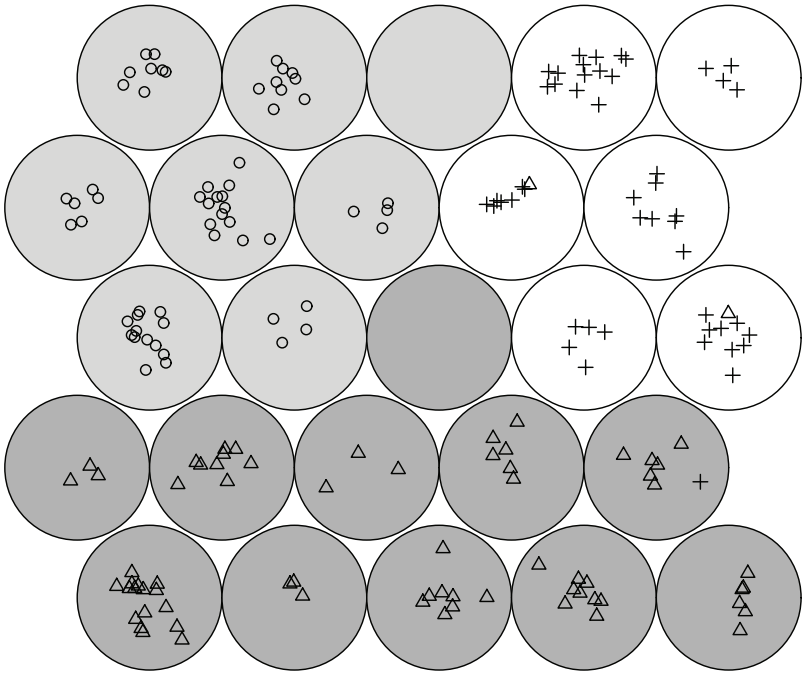

FIGURE 1.5: Self Organizing Maps. A 5 by 5 hexagonal SOM has been trained on a dataset of wines. Each point (triangle, circle, or cross) represents a wine which originally is described by a 13-dimensional vector. The shape of the point represents the category of the wine, the shade of grey of each activation unit (the big circles of the grid) is the predicted category. Wines that are similar in the 13-dimensional space are close to each other on this grid.

neighbouring branches which, in this particular research area, are very useful to study gene transfers in DNA.

1.2.5.2 Generative Topographic Mapping

Instead looking for spatial relations, like SOMs, one may think of correlations between the variables of the dataset. One way to capture this hidden structure is to model the distribution of the data in terms of hidden variables. An example of this approach is factor analysis, which is based on a linear transformation from data space to latent space. In [12], the authors extend this concept of a hidden variable framework into a Generative Topographic Mapping (GTM). The idea is very similar to the SOMs; however, the most significant difference between the GTM and SOM algorithms is that GTM defines an explicit probability density given by the mixture distribution of variables. As a consequence, there is a well-defined objective function given by the log likelihood, and convergence to a local maximum of the objective function is guaranteed by the use of the Expectation Maximization algorithm.

In [4], GTM is to cluster motor unit action potentials for the analysis of the behavior

of the neuromuscular system. The aim of the analysis is to reveal how many motor units are active during a muscle contraction. This work compares the strength and weaknesses of GTM and principal component analysis (PCA), an alternative multidimensional projection technique. The advantage of PCA is that the method allows the visualization of objects in a Euclidian space where the perception of distance is easy to understand. On the other hand, the main advantage of the GTM is that each unit may be considered as an individual cluster, and the access to these micro-clusters may be very useful for elimination or selection of wanted or unwanted information.

1.2.6 Trees

During the 1980s, the appeal of graphical user interfaces encouraged many developers to create node-link diagrams. By the early 1990s, several research groups developed innovative methods of tree browsing that offered different overview and browsing strategies. For a history of the development of visualisation tools based on trees refer to [74]. In this section, we present four variants of visualisation of trees: decision trees, tree maps, hyperbolic trees, and phylogenetic trees.

1.2.6.1 Decision Trees

A decision tree, also known as classification tree or regression tree, is a technique for partitioning data into homogeneous groups. It is constructed by iteratively partitioning the data into disjoint subsets, and one class is assigned to each leave of the tree. One of the first methods for building decision trees was CHAID [43]. This method partitions the data into mutually exclusive, and exhaustive, subsets that best describe the dependent variables.

In Figure 1.6, an example of a decision tree is shown. The dataset contains information about cars taken from the April, 1990 issue of Consumer Reports. [8] Each node of the tree predicts the average car mileage given the price, the country, the reliability, and the car type. [9] In this example, given the price and the type of the car, we are able to classify the car in different categories of gas consumption by following a path from the root to a leaf.

Decision tree visualization and exploration is important for two reasons: (i) it is crucial to be able to navigate through the decision tree to find nodes that need to be further partitioned; (ii) exploration of the decision tree aids the understanding of the tree and the data being classified. In [5], the authors present an approach to support interactive decision tree construction. They show a method for visualising multi-dimensional data with a class label such that the degree of impurity of each node with respect to class membership can be easily perceived by users. In [56], a conceptual model of the visualization support to the data mining process is proposed, together with a novel visualisation of decision tree classification process with the aim of exploring humans pattern recognition ability and domain knowledge to facilitate the knowledge discovery process. PaintingClass is a different interactive approach where the user interactively edits projections of multi-dimensional data and "paints" regions to build a decision tree [80]. The visual interaction of this systems combines Parallel Coordinates and Star Coordinates by showing this 'dual' projection of the data.

1.2.6.2 Treemap

The Treemap visualization technique [73] makes use of the area available on the display, mapping hierarchies onto a rectangular region in a space-filling manner. This efficient use

[8] `http://stat.ethz.ch/R-manual/R-devel/library/rpart/html/cu.summary.html`
[9] `http://cran.r-project.org/web/packages/rpart/`

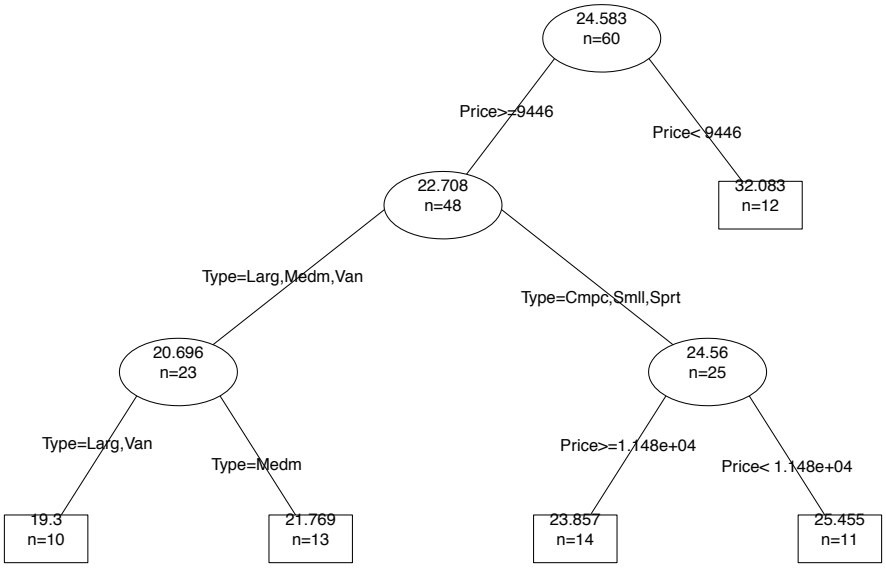

FIGURE 1.6: Decision Trees. Each node of the tree predicts the average car mileage given the price, the country, the reliability, and the car type according to the data of from April, 1990 issue of Consumer Reports. In this example, given the price and the type of the car, we are able to classify the car in different categories of gas consumption.

of space allows large hierarchies to be displayed and facilitates the presentation of semantic information. Each node of a tree map has a weight which is used to determine the size of a nodes bounding box. The weight may represent a single domain property, or a combination of domain properties. A nodes weight determines its display size and can be thought of as a measure of importance or degree of interest [40].

In Figure 1.7, a tree map shows the gross national income per country. Each box (the node of the tree) represents a country, the size of the box is proportional to the size of the population of that country. The shade of grey of the box reflects the gross national income of the year 2010. [10]

Treemaps can also displayed in 3D [30]. For example, patent classification systems intellectually organize the huge number of patents into pre-defined technology classes. To visualize the distribution of one or more patent portfolios, an interactive 3D treemap can be generated, in which the third dimension represents the number of patents associated with a category.

1.2.6.3 Hyperbolic Tree

Hyperbolic geometry provides an elegant solution to the problem of providing a focus and context display for large hierarchies [52]. The hyperbolic plane has the room to layout large hierarchies, with a context that includes as many nodes as are included by 3D approaches and with modest computational requirements. The root node in the center with first-level nodes arranged around it in a circle or oval. Further levels are placed in larger concentric circles or ovals,thus preserving a two-dimensional planar approach. To ensure that the entire tree is visible, outer levels are shrunk according to a hyperbolic formula. In [37], hyperbolic

[10] `http://cran.r-project.org/web/packages/treemap/index.html`

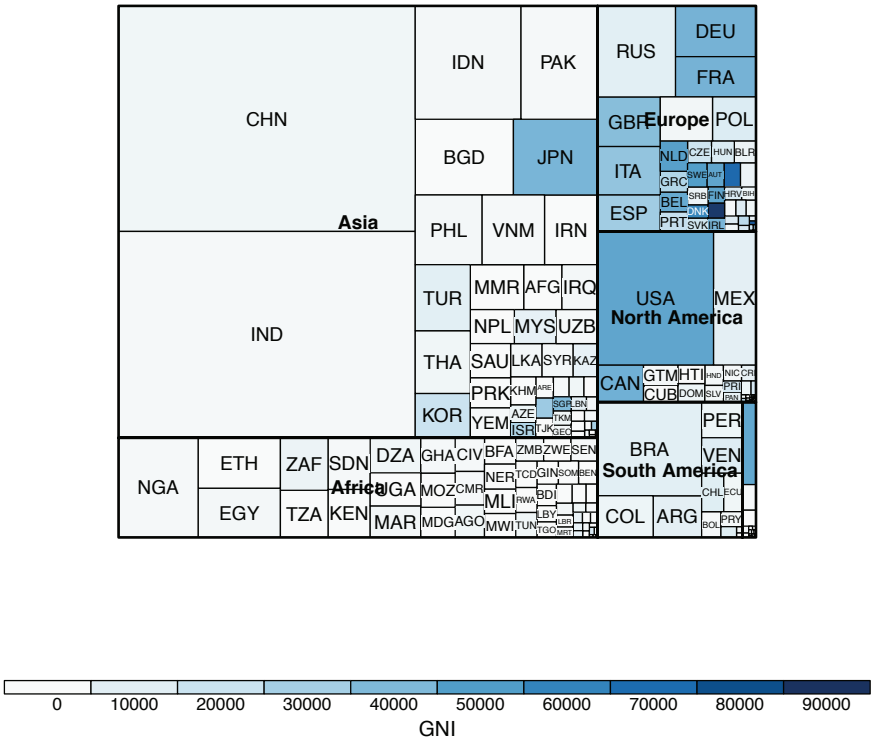

FIGURE 1.7: Treemaps. This plot represents a dataset of 2010 about population size and gross national income for each country. The size of each node of the treemap is proportional to the size of the population, while the shade of grey of each box represent the gross national income of that country. The countries of a continent are grouped together into a rectangular area.

trees are used for spam classification. The authors propose a Factors Hyperbolic Tree based algorithm that, unlike the classical word and lexical matching algorithms, handles spam filtering in a dynamic environment by considering various relevant factors.

1.2.6.4 Phylogenetic Trees

Phylogenetic trees are an alternative approach for the construction of object maps targeted at reflecting similarity relationships [17]. By means of a distance matrix, the aim is inferring ancestors for a group of objects and reconstructing the evolutionary history of each object. The main advantages of the approach are improved exploration and more clear visualization of similarity relationships, since it is possible to build an ancestry relationships from higher to lower content correlation. In [65], the authors present a phylogenetic tree to support image and text classification. They discuss some challenges and advantages for using this type of visualization. A set of visualisation tools for visual mining of images and text is made possible by the properties offered by these trees complemented by the possibilities offered by multidimensional projections.

1.3 Systems

One of the most important characteristic of a visual classification system is that users should gain insights about the data [15]. For example, how much the data within each class varies, which classes are close to or distinct from each other, see which features in the data play an important role to discriminate one class from another, and so on. In addition, the analysis of misclassified data should provide a better understanding of which type of classes are difficult to classify. Such insight can then be fed back to the classification process in both the training and the test phases.

In this section, we present a short but meaningful list of visual classification systems that have been published in the last five years and that fulfil most of the previous characteristics. The aim of this list is to address how visual classification systems support automated classification for real-world problems.

1.3.1 EnsembleMatrix and ManiMatrix

EnsembleMatrix and ManiMatrix are two interactive visualisation systems that allow users to browse and learn properties of classifiers by comparison and contrast and build ensemble classification systems. These systems are specifically designed for Human and Computer Interaction researchers who could benefit greatly from the ability to express user preferences about how a classifier should work.

EnsembleMatrix [77] allows users to create an ensemble classification system by discovering appropriate combination strategies. This system supplies a visual summary that spans multiple classifiers and helps users understand the models' various complimentary properties. EsnembleMatrix provides two basic mechanisms to explore combinations of classifiers: (i) partitioning, which divides the class space into multiple partitions; (ii) arbitrary linear combinations of the classifiers for each of these partitions.

The ManiMatrix (Manipulable Matrix) system is an interactive system that enables researchers to intuitively refine the behavior of classification systems [42]. ManiMatrix focuses on the manual refinement on sets of thresholds that are used to translate the probabilistic output of classifiers into classification decisions. By appropriately setting such parameters as the costs of misclassification of items, it is possible to modify the behavior of the algorithm such that it is best aligned with the desired performance of the system. ManiMatrix enables its users to directly interact with a confusion matrix and to view the implications of incremental changes to the matrix via a realtime interactive cycle of reclassification and visualization.

1.3.2 Systematic Mapping

Systematic mapping provides mechanisms to identify and aggregate research evidence and knowledge about when, how, and in what context technologies, processes, methods or tools are more appropriate for software engineering practices. [27] proposes an approach, named Systematic Mapping based on Visual Text Mining (SM-VTM), that applies VTM to support the categorization and classification in the systematic mapping.

The authors present two different views for systematic mapping: cluster view and chronological view. Users can explore these views and interact with them, getting information to build other visual representations of a systematic map. A case study shows that there is a significant reduction of effort and time in order to conduct text categorization and classification activities in systematic mapping if compared with manual conduction. With this

approach, it is possible to achieve similar results to a completely manual approach without the need of reading the documents of the collection.

1.3.3 iVisClassifier

The iVisClassifier system [15] allows users to explore and classify data based on Linear Discriminant Analysis (LDA), a supervised reduction method. Given a high-dimensional dataset with cluster labels, LDA projects the points onto a reduced dimensional representation. This low dimensional space provides a visual overview clusters structure. LDA enables users to understand each of the reduced dimensions and how they influence the data by reconstructing the basis vector into the original data domain.

In particular, iVisClassifier interacts with all the reduced dimensions obtained by LDA through parallel coordinates and a scatter plot. By using heat maps, iVisClassifier gives an overview about clusters relationships both in the original space and in the reduced dimensional space. A case study of facial recognition shows that iVisClassifier facilitates the interpretability of the computational model. The experiments showed that iVisClassifier can efficiently support a user-driven classification process by reducing humans search space, e.g., recomputing LDA with a user-selected subset of data and mutual filtering in parallel coordinates and the scatter plot.

1.3.4 ParallelTopics

When analyzing large text corpora, questions pertaining to the relationships between topics and documents are difficult to answer with existing visualisation tools. For example, what are the characteristics of the documents based on their topical distribution? and what documents contain multiple topics at once? ParallelTopics [23] is a visual analytics system which integrates interactive visualization with probabilistic topic model for the analysis of document collections.

ParallelTopics makes use of the Parallel Coordinate metaphor to present the probabilistic distribution of a document across topics. This representation can show how many topics a document is related to and also the importance of each topic to the document of interest. ParallelTopics also supports other tasks, which are also essential to understanding a document collection, such as summarizing the document collection into major topics, and presenting how the topics evolve over time.

1.3.5 VisBricks

The VisBricks visualization approach provides a new visual representation in the form of a highly configurable framework, that is able to incorporate any existing visualization as a building block [55]. This method carries forward the idea of breaking up the inhomogeneous data into groups to form more homogeneous subsets, which can be visualized independently and thus differently.

The visualization technique embedded in each block can be tailored to different analysis tasks. This flexible representation supports many explorative and comparative tasks. In VisBricks, there are two level of analysis: the total impression of all VisBricks together gives a comprehensive high-level overview of the different groups of data, while each VisBrick independently shows the details of the group of data it represents.

1.3.6 WHIDE

The Web-based Hyperbolic Image Data Explorer (WHIDE) system is a Web visual data mining tool for the analysis of multivariate bioimages [51]. This kind of analysis spans from the analysis of the space of the molecule (i.e. sample morphology) and molecular colocation or interaction. WHIDE utilises hierarchical hyperbolic self-organizing maps (H2SOM), a variant of the SOM, in combination with Web browser technology.

WHIDE has been applied to a set of bio-images recorded to show field of view in tissue sections from a colon cancer study and we compare tissue from normal colon with tissue classified as tumour. The result of the use of WHIDE in this particular context has shown that this system efficiently reduces the complexity of the data by mapping each of the pixels to a cluster, and provides a structural basis for a sophisticated multimodal visualization, which combines topology preserving pseudo-coloring with information visualization.

1.3.7 Text Document Retrieval

In [33], the authors describe a system for the interactive creation of binary classifiers to separate a dataset of text document into relevant and non-relevant documents for improving information retrieval tasks. The problem they present is twofold: on the one hand, supervised machine learning algorithms rely on labeled data, which can be provided by domain experts; on the other hand, the optimisation of the algorithms can be done by researchers. However, it is hard to find experts both in the domain of interest and in machine learning algorithms.

Therefore, the authors compare three approaches for interactive classifier training. These approaches incorporate active learning to various degrees in order to reduce the labeling effort as well as to increase effectiveness. Interactive visualization is then used for letting users explore the status of the classifier in context of the labeled documents, as well as for judging the quality of the classifier in iterative feedback loops.

1.4 Summary and Conclusions

The exploration of large data sets is an important problem which have many complications. By means of visualisation techniques, researchers can focus and analyse patterns of data from datasets that are too complex to be handled by automated data analysis methods. Interactive visual classification has powerful implications in leveraging the intuitive abilities of the human for this kind of data mining task. This may lead to solutions which can model classification problems in a more intuitive and unrestricted way.

The 'Big Data Era' poses new challenges for visual classification since visual solution will need to scale in size, dimensionality, data types, and levels of quality. The relevant data patterns and relationships will need to be visualized on different levels of details, and with appropriate levels of data and visual abstraction.

The integration of visualization techniques with machine learning techniques is one of the many possible research paths in the future. This is confirmed by a recent workshop named "Information Visualization, Visual Data Mining and Machine Learning" [48] the aim of which was to tighten the links between the two communities in order to explore how each field can benefit from the other and how to go beyond current hybridization successes.

Bibliography

[1] Charu C. Aggarwal. Towards effective and interpretable data mining by visual interaction. *SIGKDD Explor. Newsl.*, 3(2):11–22, January 2002.

[2] Charu C. Aggarwal. On the use of human-computer interaction for projected nearest neighbor search. *Data Min. Knowl. Discov.*, 13(1):89–117, July 2006.

[3] Charu C. Aggarwal. Toward exploratory test-instance-centered diagnosis in high-dimensional classification. *IEEE Trans. on Knowl. and Data Eng.*, 19(8):1001–1015, August 2007.

[4] Adriano O. Andrade, Slawomir Nasuto, Peter Kyberd, and Catherine M. Sweeney-Reed. Generative topographic mapping applied to clustering and visualization of motor unit action potentials. *Biosystems*, 82(3):273 – 284, 2005.

[5] Mihael Ankerst, Christian Elsen, Martin Ester, and Hans-Peter Kriegel. Visual classification: an interactive approach to decision tree construction. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 392–396, New York, NY, USA, 1999. ACM.

[6] Mihael Ankerst, Martin Ester, and Hans-Peter Kriegel. Towards an effective cooperation of the user and the computer for classification. In *KDD'00*, pages 179–188, Boston, MA, USA, 2000. ACM.

[7] Muhammad Arif. Similarity-dissimilarity plot for visualization of high dimensional data in biomedical pattern classification. *J. Med. Syst.*, 36(3):1173–1181, June 2012.

[8] Barry Becker, Ron Kohavi, and Dan Sommerfield. Visualizing the simple baysian classifier. In Usama Fayyad, Georges G. Grinstein, and Andreas Wierse, editors, *Information visualization in data mining and knowledge discovery*, pages 237–249. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

[9] Barry G. Becker. Using mineset for knowledge discovery. *IEEE Computer Graphics and Applications*, 17(4):75–78, 1997.

[10] Andreas Becks, Stefan Sklorz, and Matthias Jarke. Exploring the semantic structure of technical document collections: A cooperative systems approach. In Opher Etzion and Peter Scheuermann, editors, *CoopIS*, volume 1901 of *Lecture Notes in Computer Science*, pages 120–125. Springer, 2000.

[11] Enrico Bertini and Denis Lalanne. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, VAKD '09, pages 12–20, New York, NY, USA, 2009. ACM.

[12] Christopher M. Bishop, Markus Svensén, and Christopher K. I. Williams. Gtm: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.

[13] Clifford Brunk, James Kelly, and Ron Kohavi. Mineset: An integrated system for data mining. In Daryl Pregibon David Heckerman, Heikki Mannila, editor, *KDD-97*, pages 135–138, Newport Beach, CA, USA, August 14-17 1997. AAAI Press.

[14] Matthew Chalmers and Paul Chitson. Bead: Explorations in information visualization. In Nicholas J. Belkin, Peter Ingwersen, and Annelise Mark Pejtersen, editors, *SIGIR*, pages 330–337. ACM, 1992.

[15] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 27–34, 2010.

[16] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[17] A.M. Cuadros, F.V. Paulovich, R. Minghim, and G.P. Telles. Point placement by phylogenetic trees and its application to visual analysis of document collections. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 99–106, 2007.

[18] Maria Cristina Ferreira de Oliveira and Haim Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE Trans. Vis. Comput. Graph.*, 9(3):378–394, 2003.

[19] S. Deegalla and H. Bostrom. Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *Machine Learning and Applications, 2006. ICMLA '06. 5th International Conference on*, pages 245–250, 2006.

[20] Inderjit S. Dhillon, Dharmendra S. Modha, and W.Scott Spangler. Class visualization of high-dimensional data with applications. *Computational Statistics & Data Analysis*, 41(1):59 – 90, 2002. ¡ce:title¿Matrix Computations and Statistics¡/ce:title¿.

[21] Giorgio Maria Di Nunzio. Using scatterplots to understand and improve probabilistic models for text categorization and retrieval. *Int. J. Approx. Reasoning*, 50(7):945–956, 2009.

[22] Stephan Diehl, Fabian Beck, and Michael Burch. Uncovering strengths and weaknesses of radial visualizations—an empirical approach. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):935–942, November 2010.

[23] Wenwen Dou, Xiaoyu Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 231–240, 2011.

[24] G. Draper, Y. Livnat, and R.F. Riesenfeld. A survey of radial methods for information visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 15(5):759–776, 2009.

[25] N. Elmqvist and J. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *Visualization and Computer Graphics, IEEE Transactions on*, 16(3):439–454, 2010.

[26] Daniel Engel, Lars Hüttenberger, and Bernd Hamann. A survey of dimension reduction methods for high-dimensional data analysis and visualization. In Christoph Garth, Ariane Middel, and Hans Hagen, editors, *VLUDS*, volume 27 of *OASICS*, pages 135–149. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, 2011.

[27] Katia Romero Felizardo, Elisa Yumi Nakagawa, Daniel Feitosa, Rosane Minghim, and José Carlos Maldonado. An approach based on visual text mining to support categorization and classification in the systematic mapping. In *Proceedings of the 14th international conference on Evaluation and Assessment in Software Engineering*, EASE'10, pages 34–43, Swinton, UK, UK, 2010. British Computer Society.

[28] José Roberto M. Garcia, Antônio Miguel V. Monteiro, and Rafael D. C. Santos. Visual data mining for identification of patterns and outliers in weather stations' data. In *Proceedings of the 13th international conference on Intelligent Data Engineering and Automated Learning*, IDEAL'12, pages 245–252, Berlin, Heidelberg, 2012. Springer-Verlag.

[29] Zhao Geng, ZhenMin Peng, R.S. Laramee, J.C. Roberts, and R. Walker. Angular histograms: Frequency-based visualizations for large, high dimensional data. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2572–2580, 2011.

[30] M. Giereth, H. Bosch, and T. Ertl. A 3d treemap approach for analyzing the classificatory distribution in patent portfolios. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pages 189–190, 2008.

[31] Charles D. Hansen and Chris R. Johnson. *Visualization Handbook*. Academic Press, 1 edition, December 2004.

[32] Frank Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer, second edition, Feb 2006.

[33] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2839–2848, 2012.

[34] William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors. *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*. ACM, 2012.

[35] Patrick Hoffman, Georges Grinstein, Kenneth Marx, Ivo Grosse, and Eugene Stanley. Dna visual and analytic data mining. In *Proceedings of the 8th conference on Visualization '97*, VIS '97, pages 437–ff., Los Alamitos, CA, USA, 1997. IEEE Computer Society Press.

[36] Patrick Hoffman, Georges Grinstein, and David Pinkney. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, NPIVM '99, pages 9–16, New York, NY, USA, 1999. ACM.

[37] Hailong Hou, Yan Chen, R. Beyah, and Yan-Qing Zhang. Filtering spam by using factors hyperbolic tree. In *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, pages 1–5, 2008.

[38] A. Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Visualization, 1990. Visualization '90., Proceedings of the First IEEE Conference on*, pages 361–378, 1990.

[39] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[40] Brian Johnson. Treeviz: treemap visualization of hierarchically structured information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, pages 369–370, New York, NY, USA, 1992. ACM.

[41] Eser Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In Doheon Lee, Mario Schkolnick, Foster J. Provost, and Ramakrishnan Srikant, editors, *KDD*, pages 107–116. ACM, 2001.

[42] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1343–1352, New York, NY, USA, 2010. ACM.

[43] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(2):pp. 119–127, 1980.

[44] M. W. Kattan, J. A. Eastham, A. M. Stapleton, T. M. Wheeler, and P. T. Scardino. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *J Natl Cancer Inst*, 90(10):766–71, 1998.

[45] Daniel Keim and Leishi Zhang. Solving problems with visual analytics: challenges and applications. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '11, pages 1:1–1:4, New York, NY, USA, 2011. ACM.

[46] Daniel A. Keim, Ming C. Hao, Umeshwar Dayal, Halldor Janetzko, and Peter Bak. Generalized scatter plots. *Information Visualization*, 9(4):301–311, December 2010.

[47] Daniel A. Keim, Joern Kohlhammer, Geoffrey Ellis, and Florian Mansmann, editors. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, November 2010.

[48] Daniel A. Keim, Fabrice Rossi, Thomas Seidl, Michel Verleysen, and Stefan Wrobel. Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081). *Dagstuhl Reports*, 2(2):58–83, 2012.

[49] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. In James A. Anderson and Edward Rosenfeld, editors, *Neurocomputing: foundations of research*, pages 511–521. MIT Press, Cambridge, MA, USA, 1982.

[50] Teuvo Kohonen. *Self-Organizing Maps*. Springer Series in Information Retrieval. Springer, second edition, March 1995.

[51] Jan Kölling, Daniel Langenkämper, Sylvie Abouna, Michael Khan, and Tim W. Nattkemper. Whide - a web tool for visual data mining colocation patterns in multivariate bioimages. *Bioinformatics*, 28(8):1143–1150, April 2012.

[52] John Lamping, Ramana Rao, and Peter Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 401–408, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[53] Gregor Leban, Blaz Zupan, Gaj Vidmar, and Ivan Bratko. Vizrank: Data visualization guided by machine learning. *Data Min. Knowl. Discov.*, 13(2):119–136, 2006.

[54] Ioannis Leftheriotis. Scalable interaction design for collaborative visual exploration of big data. In *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces*, ITS '12, pages 271–276, New York, NY, USA, 2012. ACM.

[55] A. Lex, H. Schulz, M. Streit, C. Partl, and D. Schmalstieg. Visbricks: Multiform visualization of large, inhomogeneous data. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2291–2300, 2011.

[56] Yan Liu and Gavriel Salvendy. Design and evaluation of visualization support to facilitate decision trees classification. *International Journal of Human-Computer Studies*, 65(2):95 – 110, 2007.

[57] H. Ltifi, M. Ben Ayed, A.M. Alimi, and S. Lepreux. Survey of information visualization techniques for exploitation in kdd. In *Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on*, pages 218–225, 2009.

[58] J. Lubsen, J. Pool, and E. van der Does. A practical device for the application of a diagnostic or prognostic function. *Methods of information in medicine*, 17(2):127–129, April 1978.

[59] Christian Martin, Naryttza N. Diaz, Jörg Ontrup, and Tim W. Nattkemper. Hyperbolic som-based clustering of dna fragment features for taxonomic visualization and classification. *Bioinformatics*, 24(14):1568–1574, July 2008.

[60] Dieter Merkl. Text classification with self-organizing maps: Some lessons learned. *Neurocomputing*, 21(1–3):61 – 77, 1998.

[61] Martin Mozina, Janez Demsar, Michael W. Kattan, and Blaz Zupan. Nomograms for visualization of naive bayesian classifier. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *PKDD*, volume 3202 of *Lecture Notes in Computer Science*, pages 337–348. Springer, 2004.

[62] Emmanuel Müller, Ira Assent, Ralph Krieger, Timm Jansen, and Thomas Seidl. Morpheus: interactive exploration of subspace clustering. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 1089–1092, New York, NY, USA, 2008. ACM.

[63] Giorgio Maria Di Nunzio and Alessandro Sordoni. A visual tool for bayesian data analysis: the impact of smoothing on naive bayes text classifiers. In Hersh et al. [34], page 1002.

[64] P. Oesterling, G. Scheuermann, S. Teresniak, G. Heyer, S. Koch, T. Ertl, and G.H. Weber. Two-stage framework for a topology-based projection and visualization of classified document collections. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 91–98, 2010.

[65] J.G. Paiva, L. Florian, H. Pedrini, G.P. Telles, and R. Minghim. Improved similarity trees and their application to visual data classification. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2459–2468, 2011.

[66] François Poulet. Towards effective visual data mining with cooperative approaches. In Simoff et al. [75], pages 389–406.

[67] Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russell Greiner, David S. Wishart, Alona Fyshe, Brandon Pearcy, Cam Macdonell, and John Anvik. Visual explanation of evidence with additive classifiers. In *AAAI*, pages 1822–1829. AAAI Press, 2006.

[68] Bhiksha Raj and Rita Singh. Classifier-based non-linear projection for adaptive endpointing of continuous speech. *Computer Speech & Language*, 17(1):5–26, 2003.

[69] Randall M. Rohrer, John L. Sibert, and David S. Ebert. A shape-based visual interface for text retrieval. *IEEE Computer Graphics and Applications*, 19(5):40–46, 1999.

[70] Christin Seifert and Elisabeth Lex. A novel visualization approach for data-mining-related classification. In Ebad Banissi, Liz J. Stuart, Theodor G. Wyeld, Mikael Jern, Gennady L. Andrienko, Nasrullah Memon, Reda Alhajj, Remo Aslak Burkhard, Georges G. Grinstein, Dennis P. Groth, Anna Ursyn, Jimmy Johansson, Camilla Forsell, Urska Cvek, Marjan Trutschl, Francis T. Marchese, Carsten Maple, Andrew J. Cowell, and Andrew Vande Moere, editors, *IV*, pages 490–495. IEEE Computer Society, 2009.

[71] B. Shneiderman. Direct manipulation: A step beyond programming languages. *Computer*, 16(8):57–69, 1983.

[72] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343, 1996.

[73] Ben Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.*, 11(1):92–99, January 1992.

[74] Ben Shneiderman, Cody Dunne, Puneet Sharma, and Ping Wang. Innovation trajectories for information visualizations: Comparing treemaps, cone trees, and hyperbolic trees. *Information Visualization*, 11(2):87–105, 2012.

[75] Simeon J. Simoff, Michael H. Böhlen, and Arturas Mazeika, editors. *Visual Data Mining - Theory, Techniques and Tools for Visual Analytics*, volume 4404 of *Lecture Notes in Computer Science*. Springer, 2008.

[76] Rita Singh and Bhiksha Raj. Classification in likelihood spaces. *Technometrics*, 46(3):318–329, 2004.

[77] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S. Tan. Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1283–1292, New York, NY, USA, 2009. ACM.

[78] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, and D. Keim. Automated analytical methods to support visual exploration of high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on*, 17(5):584–597, 2011.

[79] Soon T. Teoh. StarClass: Interactive Visual Classification Using Star Coordinates - CiteSeerX. *Proceedings of the 3rd SIAM International Conference on Data Mining*, 2003.

[80] Soon Tee Teoh and Kwan-Liu Ma. Paintingclass: interactive construction, visualization and exploration of decision trees. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 667–672, New York, NY, USA, 2003. ACM.

[81] Clark T.G., Stewart M.E., Altman D.G., Gabra H., and Smyth J.F. A prognostic model for ovarian cancer. *British Journal of Cancer*, 85(7):944–952, October 2001.

[82] Michail Vlachos, Carlotta Domeniconi, Dimitrios Gunopulos, George Kollios, and Nick Koudas. Non-linear dimensionality reduction techniques for classification and visualization. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 645–651, New York, NY, USA, 2002. ACM.

[83] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H. Witten. Interactive machine learning: letting users build classifiers. *Int. J. Hum.-Comput. Stud.*, 56(3):281–292, March 2002.

[84] James A. Wise, James J. Thomas, Kelly Pennock, D. Lantrip, M. Pottier, Anne Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In Nahum D. Gershon and Stephen G. Eick, editors, *INFOVIS*, pages 51–58. IEEE Computer Society, 1995.

[85] Pak Chung Wong. Guest editor's introduction: Visual data mining. *IEEE Computer Graphics and Applications*, 19(5):20–21, 1999.

[86] Jing Yang, Wei Peng, Matthew O. Ward, and Elke A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of the Ninth annual IEEE conference on Information visualization*, INFOVIS'03, pages 105–112, Washington, DC, USA, 2003. IEEE Computer Society.

[87] Ke-Bing Zhang, M.A. Orgun, R. Shankaran, and Du Zhang. Interactive visual classification of multivariate data. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, pages 246–251, 2012.

[88] Ke-Bing Zhang, M.A. Orgun, and Kang Zhang. A visual approach for external cluster validation. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*, pages 576–582, 2007.

[89] Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui, and Baoquan Chen. Visual clustering in parallel coordinates. *Computer Graphics Forum*, 27(3):1047–1054, 2008.