Frequency Plot and Relevance Plot to Enhance Visual Data Exploration

José Fernando Rodrigues Jr., Agma J. M. Traina, Caetano Traina Jr.
Computer Science Department
University of Sao Paulo at Sao Carlos - Brazil
Avenida Trabalhador Sãocarlense, 400
13.566-590 São Carlos, SP - Brazil
e-mail: [junio, cesar, caetano, agma]@icmc.usp.br

Abstract

This paper presents two techniques aiming at exploring databases through multivariate visualizations. Both techniques intend to deal with the problem caused by the limited amount of elements that can be presented simultaneously in traditional visual exploration procedures. The first technique, the Frequency Plot, combines data frequency with interactive filtering to identify clusters and trends in subsets of the database. Thus, graphical elements (lines, pixels, icons, or graphical marks) are color differentiated proportionally to how frequent the value being represented is, while interactive filtering allows the selection of interesting partitions of the database. The second technique presented in this work, the Relevance Plot, corresponds to assigning different levels of color distinguishably to visual elements according to their relevance to a user's specified data properties set, which can be chosen visually and dynamically.

1. Introduction

The volume of digital data generated by the enterprises around the world has increased exponentially in the last years. Therefore, together with concerning about efficient storage and fast and effective retrieval of the information, come the concern of getting the right information at the right time. That is, companies can gain marketing space by knowing more about their clients' preferences, usual transactions, and trends. Well-organized information can be a valuable and strategic asset, providing competitive advantage on business activities.

As the amount of the data is usually huge, what sometimes happens is the process of finding a "needle in a haystack". Information Visualization (*Infovis*) techniques aim at helping the human beings to absorb the inherent behavior of the data and to easily recognize relationships among the information elements. Therefore, such techniques are becoming more and more important during data exploration and analysis. Information visualization techniques take advantage of the fact that humans can understand much more easily and faster through graphical presentations.

Besides the increasing amount of data that the information systems have to deal with, the great majority of them manage multidimensional information. The process of analyzing these data is cumbersome, because this type of information is complex, as it is constituted by many features and properties that must be controlled. The dimensions (attributes) of multi-dimensional data can be seen as points in some *k*-dimensional space, where *k* is the number of dimensions. Thus, Information Visualization techniques intend to map a database with elements in a *k*-dimensional space into the two dimensions of the computer display.

Our first proposed technique, the *Frequency Plot*, intends to tackle with two problems, both derived from the increasing in the amount of data, observed in most of the known visualization techniques in literature. These two concerning issues are the overlapping of graphical elements, and the excessive population of the visualization scene. The former prevents the techniques from presenting information implicit in the "lost" elements that coincided in the scene. The later is responsible for determining unintelligible visualizations since, due to the over population, no tendency or evidence can be perceived.

These cases are exemplified in figure 1 through the use of the Parallel Coordinates technique [1]. Figure 1(a) shows a common database where some ranges are so massively populated that only blots can be seen in the visualization scene. Therefore, the hidden elements cannot contribute for investigation. In figure 1(b), it is shown a hypothetical

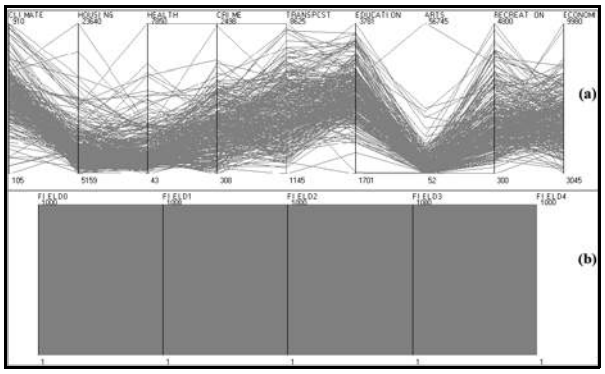

Figure 1 Examples of databases with too concentrated data (a), and with too spread data (b).

database where all dimensions are restricted to the discrete integer domain, and where all the values between the lowest and the highest limits fully populate every dimension, thus,the correspondent visualization becomes simply a meaningless mass.

Therefore, aiming at dealing with the issues pointed out above, the *Frequency Plot* intends to increase the analytical power of visualization techniques of all natures by providing a method through which data can be selectively analyzed. The analysis, based on data frequency, can demonstrate the areas where the database is most populated, while its interactive characteristic allows the user to choose the most interesting subsets where this analysis should take place.

Our second proposed technique, the *Relevance Plot*, describes a way to verify and to present, the behavior of a data set based on a user's interactively defined set of properties, over which the records will be confronted in order to determine their importance according to what was stated as important by the user, that is, to what is more relevant to the analyst. This procedure permits the verification, the discovering and the validation of hypothesis, providing a data insight capable of revealing essential features of the data.

The remainder of this paper is structured as follows. Section 2 describes the dataset used in the tests presented in this work. Section 3 gives an overview of the related literature used as reference for the current proposes. Section 4 and section 5 present the *Frequency Plot* and *Relevance Plot* approaches respectively. Finally, the conclusions and future work are discussed in section 6.

2. The breast cancer dataset

The breast cancer data set [2], used in our experiments, was built by Dr. William H. Wolberg and Olvi Mangasarian from the University of Wisconsin. It comprises 457 records from patients whose identities were removed. Each data item is described by 11 attributes, including a numeric sample identifier (attribute 0) and a classifier (attribute 11) which indicates the tumor occurrence type (0 for benign and 1 for malign). The remaining fields are data originated from analytical tests over patients' tissue samples carried on clinical laboratories. The tests might indicate the malignity degree of breast cancer. The other attribute names, from the 2nd to the 10th, are *ClumpThickness, UniforSize, UniforShape, MargAdhes, SingleEpithSize, BareNuclei, BlandChromatin, NormalNucleoli and Mitoses.* These names have meaning restricted to medical domain. Noise data was removed from the original source.

3. Background and Related Work

According to [3], conventional multivariate visualization techniques do not scale well with respect to the number of objects in the data set, resulting in a display with an unacceptable level of clutter. In [4] it is affirmed that the maximum number of elements that the Parallel Coordinates technique can present is around one thousand. In fact, most of the visualization techniques do not comprise with much more than that number, either due to space limitations inherent to current display devices, or due to data sets whose elements tend to spread over the data domain. Therefore, such visualizations give scenes with a reduced number of noticeable differences.

Greatly populated databases inevitably have overlapping values, or a too spread distribution of data, what degenerates many multivariate visualization techniques as the Parallel Coordinates and Scatter Plots [5]. These shortcomings have been dealt by the computer science community in many works on the area, as follows.

A very efficient method to bypass the limitations of overploting is using hierarchical clustering to generate the visualization while expressing aggregation information. The work in [6] proposes a complete navigation system to allow the user toachieve the intended level of details in the areas of interest. The proposal is initially schematized for the use with the Parallel Coordinates, but an implementation that comprises many other multivariate visualization schemes is available in the Xmdv Tool [3]. The drawback of this system is its complex navigating interface and the need for high processing power for constant clustering of the data based on user redefinitions.

Another approach worth to mention is presented in [7], which uses wavelets to present data in lower resolutions without losing its original behavior. This technique takes

advantage of the wavelets' intrinsic property of image details reduction. Although there is a predicted data loss that might degrade the analysis capabilities, the use of this tool can enhance the dynamic filtering activity.

However, the most known and used alternative to troubleshoot the anomalies of massive datasets is to find means to visually present subsets of the data instead of the whole database in a single scene. The interactive filtering principle, according to [8], claims that "in exploring large data sets, it is important to interactively partition the data set into segments and focus on interesting subsets". Following that principle, many other authors developed tools aiming the interactive filtering goal, as the Magic Lenses [9] and the Dynamic Queries [10]. This selective visualization is fundamental in interaction mechanisms since it enriches the user participation during the visualization process, allowing users to focus on partitions of more interest by constantly redefining the scene in order to find more relevant information that will characterize the data under analysis.

An interesting approach in selective exploration is presented in the VisdB tool [11], which determines a complete interface to specify a query whose results will be the basis for the resulting scene. The presentation of data takes advantage of their relevance to color information items. A color scheme determines the hue of the visual items according to their proximity to the items returned by the query. The multivariate technique utilized is pixel oriented and the interaction depends on a query form positioned aside the visualized query result. Also, the analysis depends on the user capability to join information originating from one window per dimension, since each dimension visualized is docked in a separate scene.

The last topic to be reviewed as a basis for our development is the direct manipulation applied to Information Visualization [12], This might be understood as the ability of the user to interact with a visualization scene in such a way that the reaction of the system to the user's activity occurs within an amount of time short enough for the user to establish a correlation between what happened in the scene and his/her action[13]. Known as the cause and effect time limit[12], this time is approximately 0.1 second, the maximum accepted time before action and reaction seems disjoining. This concept is essential since no implementation is worth for the user if this human-computer interaction principle is not respected.

4. The GBDIView Tool

In order to apply and test our ideas, we implemented a tool,

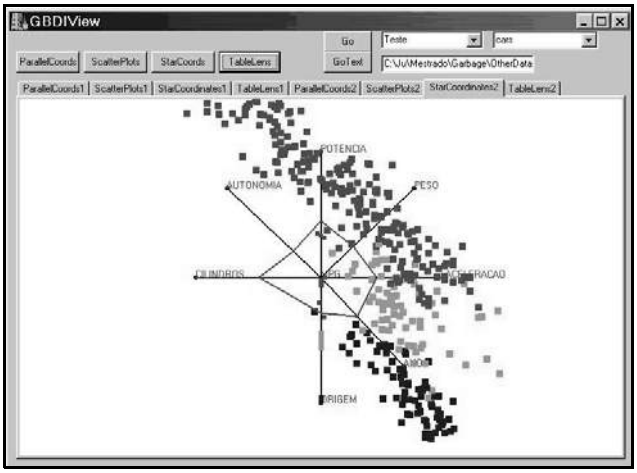

Figure 2 The GBDIView tool presenting a Star Coordinates visualization with interactive filtering.

presented in figure 2, that fully comprises the theory presented above and extends it. The GBDIView tool is an application that consists of 4 well-known visualization techniques enhanced by the proposed techniques we have developed. The tool, built in C++, was designed following the software reuse paradigm, therefore, being idealized as a set of visualization techniques in the form of software components totally adaptable to any application that might be conceived in the field of Infoviz.

The techniques embodied by the tool are the Parallel Coordinates, the Scatter Plots Matrix, the Star Coordinates [14] and the Table Lens [15]. The four visual schemes are integrated by the Link & Brush [16] technique and are also enabled with statistics data analysis (average, standard deviation, median), which can be presented graphically over the rendered scenes.

A fully functional version of the tool, along with sample data sets and user's guide, can be downloaded at http://gbdi.icmc.usp.br/~junio/GBDIViewTool.htm.

5. The Frequency Plot with Interactive Filtering

Now we present a new enhancement for Information Visualization techniques, which intend to, by one side, bypass the limits of visualization techniques previously pointed out in this text and, by the other side, propose a manner to multiply the analytical power of every kind of technique through a composed interaction mechanism. It combines interactive filtering with direct manipulation in an analytical scheme of presentation. That is, selective visualization with the enhancement that the filtering is followed by automatic analysis of the selected portions of

the data set. We named this presentation as *Frequency Plot*.

Here we describe the idea in general terms, reserving an example for later detailing. By frequency, we mean how common (frequent) a determined value can be found inside a set of values. Formally:

Given a set of values $V = \{v_0, v_1, ..., v_{k-1}\}$, let be a function $q(v,V) \rightarrow N$, which counts how many times $v \in V$ appears inside the set V. Also given a function m(V), that returns the statistical mode value of the set V. The frequency coefficient of a value $v \in V$ is given by:

$$f(v,V) = q(v,V) / q(m(V),V) \quad \textbf{(1)}$$

The function $f(v,V)$ returns a real number between 0 and 1 that indicates how frequently the value v is found inside the set V. In our work, this function is applied for every value of every dimension of the range under analysis. Given a dataset C with n elements and k dimensions, its values might be faced as a set D with k subsets of values, a subset for each data dimension, that is, $D = \{\{D^0\},\{D^1\},...,\{D^{k-1}\}\}$, having $|D^x| = n$. Given a k-dimensional data item $c_j = (c_j^0, c_j^1, ..., c_j^{k-1})$ that belongs to the set C, its correspondent k-dimensional vector of frequencies F_j is given by:

$$F_j = \left(f\left(c_j^0, D^0\right), f\left(c_j^1, D^1\right), ..., f\left(c_j^{k-1}, D^{k-1}\right)\right) \quad \textbf{(2)}$$

Through the use of the function presented in Equation 2, we can calculate the frequency data for every k-dimensional element. And, once calculated the frequencies, the idea is to exhibit them through visual effects as color and size. This is demonstrated on our implementation of the Parallel Coordinates technique, where the frequencies are expressed by color, and on our implementation of the Scatter Plots technique, where they are expressed by both color and size. In our project, based on a single color for data items and on a white background for scene, the high density values were exhibited with more saturated tones, in contrast with the low density ones, whose visualization was based on less visible graphical elements, since smooth saturations tend to disappear in white background.

Coupling the interactive filtering with this idea, our proposed visualization is not based on the whole data set, but on partitions of it that are specified by the user. These partitions must be acquired through the manipulation stated by interactive filtering principles, embodying logical operators being the user able to select data through logic combinations of dimension ranges, named AND, OR and XOR. Therefore, only the data items that satisfy the visual queries are used to perform the frequency analysis and, thus, subsets of the database can be characterized to better demonstrate data properties.

The direct manipulated query allows the user to focus on specific subsets selected visually, while the generated frequency visualization instantly characterizes this limited bundle of elements. Tendencies and the main axes of the

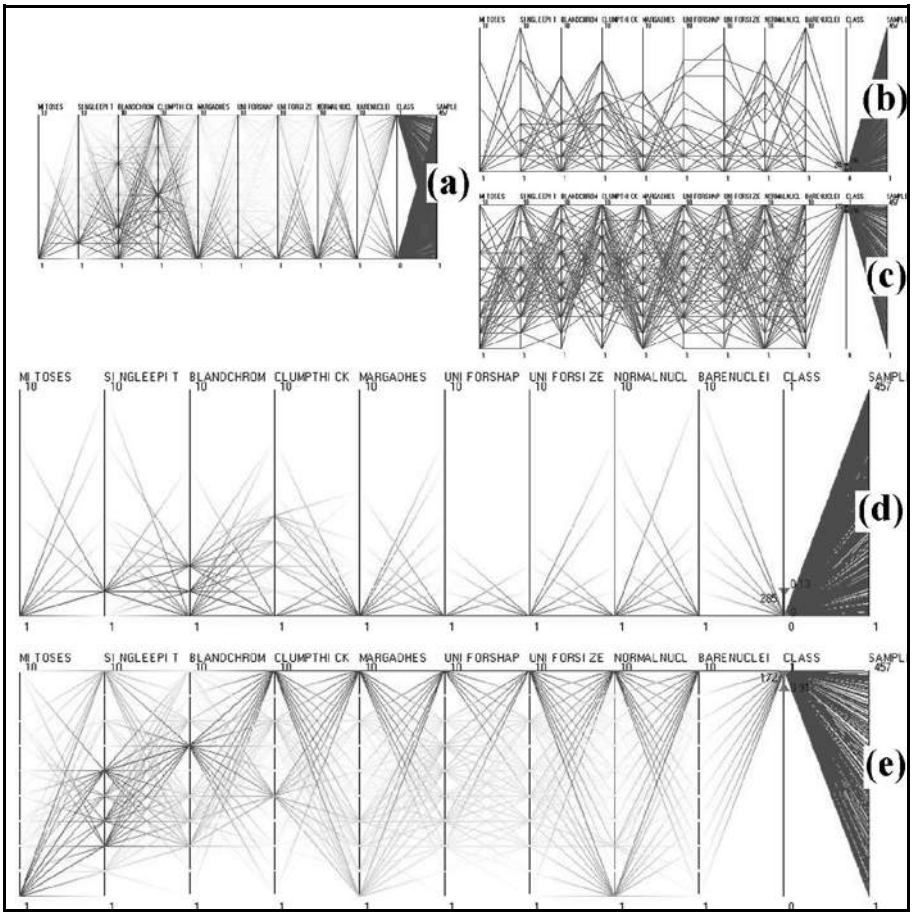

Figure 3 Parallel Coordinates with Frequency Plot. The frequency analysis of the whole dataset (a), the traditional visualization of class 1 (malign) (b) and class 0 (benign) (c) records. And in (d) and (e) the correspondent Frequency Plot of these classes.

selected data come up in scene and conclusions can be immediately produced. The high-density and more visible areas of the scene indicate local clusters, while the low-density thinner ones indicate outliers that might reveal special conditions, as data incoherence or exceptions. Also, departing from the fact that all the procedure started with an analyst-provided filter, the investigation occurs in an optimized fashion that generates precious results in less time.

An illustrative example might be seen in figure 3, where the *Frequency Plot* of a complete test data set and the traditional range query approach are being contrasted with a implementation that comprises range query with *Frequency plot.* The dataset under analysis is the breast cancer autopsy database described in section 2. There is illustrated an analysis process intended to clarify what is the difference between malign and benign breast cancer based on laboratory tests.

In figure 3(a), the overall data distribution can be observed through the use of a global frequency analysis. The image indicates more presence of lower values in mostly dimensions. In figures 3(b) and 3(c) the malign and benign records are presented through the use of ordinary filtering, which simple color differentiates the data items. It becomes clear that these three scenes little contribute to cancer characterization, as the visualization should do. None of them can partition and analyze data simultaneously and, consequently, are incapable of supporting a consistent parsing of the problem.

In contrast, figures 3(d) and 3(e) demonstrate the cancer main characteristics relative to the laboratory tests and cancer nature. By highlighting the most populated areas of the selections, malign and benign cancer turns to be easily identified by searching for patterns alike those that are made explicit by the *Frequency Plot*. Therefore, the analyst is enabled to conclude what results to search for in order to help on the characterization of the cancer nature. The data distribution is visually noticed and can be separately appreciated thanks to interactive filtering.

Figure 4 shows the same visualizations as were presented in figure 3, but in a Scatter Plots Matrix enhanced by the *Frequency Plot* analysis. The Scatter Plots visualization corroborates what has been seen in the Parallel Coordinates scenes and it also clarifies the fact that the "*BlandChromation*" and the "*SingleEpithSize*" attributes of the data set are the less categorical ones and, therefore, the less discriminative in cancer nature classification.

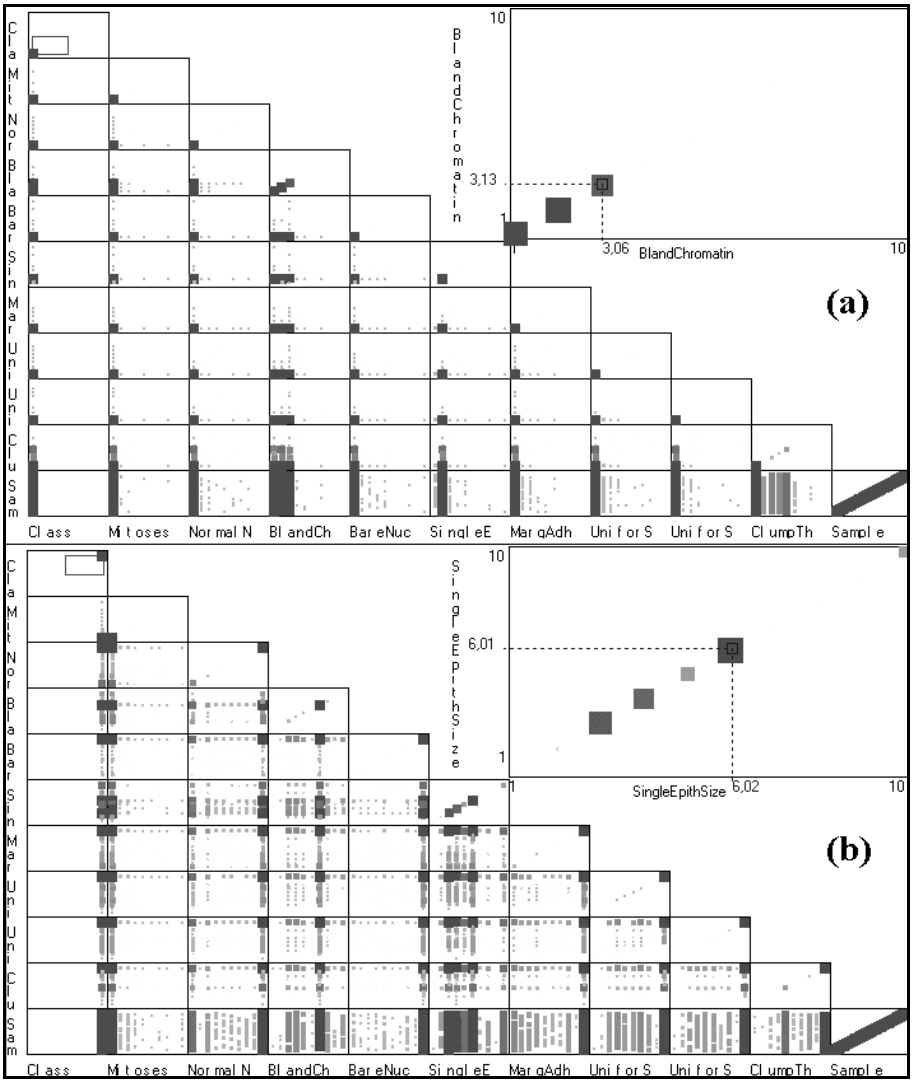

Figure 4 The Scatter Plot Matrix with *Frequency Plot* showing the benign cancer autopsies in (a) and the malign cancer autopsies in (b). Also, in (a) it can be observed the zoom of the *"BlandChromation"* attribute, and in (b) the zoom of the *"SingleEpithSize"* attribute.

The use of queries in such a way to determine the subset of interest for frequency plotting is very powerful. Now cluster identification is not limited to the whole data set, nor to predefined manipulated data records, instead, it might occur in function of a single data value. One might choose a dense populated area belonging to one of the dimensions and question the reason for that behavior. The frequency plotting will immediately present the most frequent values of the other dimensions that are correlated to the region of interest; correlation goes straight along with partial cluster identification.

6. The Relevance Plot

The second technique described in this work is based on the concept of data relevance to show the information according to the user's needs and sensibility. Therefore, we intend to reduce the amount of information presented by drawing data items using patterns in accordance to their relevance to the analysis. That is, if the data has a strong impact on the information understanding, their visualization ought to stress this fact, and the opposite must happen to data that is not relevant for the user grasping of information. Pursuing this goal, the relevance plot, described here, benefits from computer graphic facilities to depict automated data analysis through color, size, position and selective brightening.

The interaction we defend is not dependent of a query stated on Structured Query Language (SQL). Thus, it is not based on a set of data ranges, but rather, it is based on a set of data values considered interesting. These values that belong to the database dimensions' domains are used to determine how relevant each data item is. Once these relevant characteristics are set, automated analysis proceeds by calculating data relevance relative to what was chosen to be more interesting.

The mechanism, exemplified in figure 5, requires that the analyst chooses values, or *Relevance Points*, from the dimensions being visualized. Hence, given a set of data items *C* with *n* elements and *k* dimensions, assumed to be previously normalized so each dimension ranges from 0.0 to 1.0, the following definitions hold:

Definition 1: the *Relevance Point* (*RP*) of the *i*-th dimension, or RP^i, is the chosen **value belonging to the *i*-th dimension domain** that must be considered to determine the data relevance in that dimension. Only one *RP* might be chosen per dimension.

Once the *Relevance Points* are set, the data items belonging to the database must be analyzed relatively to these points. So, in each of the dimensions that had a chosen *RP*, all of the values (attributes) have computed their Euclidean distance to the respective relevance value.

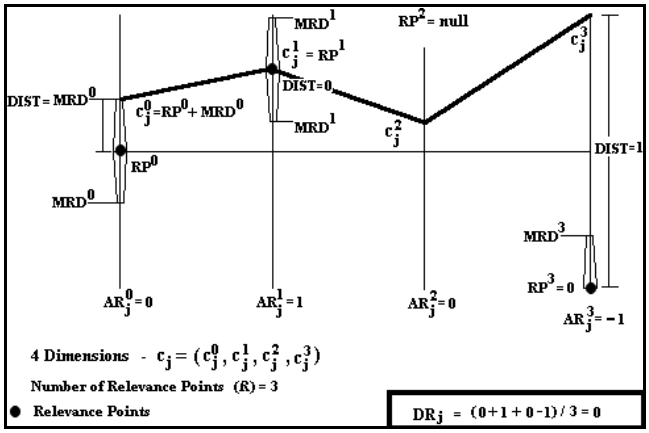

Figure 5 The *Relevance Plot* schema is demonstrated here through the calculus of the relevance for a 4-dimensional sample record visualized in the Parallel Coordinates technique.

Definition 2: for the *j*-th k dimensional data record $c_j = (c_j^0, c_j^1, ..., c_j^{k-1})$, the distance of its *i*-th attribute to the *i*-th *RP*, or $D_j^i(c_j^i, RP^i)$, is given by:

$$D_j^i(c_j^i, RP^i) = |c_j^i - RP^i| \qquad (3)$$

Also, for each of the dimensions of the *k*-dimensional database, a maximum acceptance distance is defined. These thresholds are called *Max Relevance Distance*s, or *MRDs* and are used in the relevance analysis.

Definition 3: the *Max Relevance Distance* of the *i*-th dimension, or MRD^i, is the maximum distance $D_j^i(c_j^i, RP^i)$ a data attribute can assume, before having its relevance decreased during relevance analysis. The *MRDs* take values within the range [0.0, 1.0].

Based on the *MRDs* and on the calculated distances $D_j^i(c_j^i, RP^i)$, a value named *Attribute Relevance* (*AR*) is computed for each attribute of the *k*-dimensional data records. Thus, a total of *k ARs* are computed for each of the *n k*-dimensional data records of the database.

Definition 4: the value that determines the contribution of the *i*-th attribute of the *j*-th data item, c_j^i, in the relevance analysis is called *Attribute Relevance*, and is given by:

$$AR_j^i = \begin{cases} 1 - \dfrac{(D_j^i(c_j^i, RP^i))}{MRD^i} & \text{if} \quad D_j^i(c_j^i, RP^i) \le MRD^i \\ -\dfrac{(D_j^i(c_j^i, RP^i))}{(1 - MRD^i)} & \text{if} \quad D_j^i(c_j^i, RP^i) > MRD^i \\ 0 & \text{if} \quad RP^i = -1 \end{cases} \qquad (4)$$

Equation 4 states that:

• For distances *D(c,RP)* smaller or equal the *MRD*, the equation has been settled to assign values ranging from 1 (where the distances *D*(*c*,*RP*) are null) to 0 (for distances equal the *MRD*);

• For distances *D*(*c*,*RP*) bigger than the *MRD*, the equation linearly assigns values ranging from 0 to -1, this last value is assigned to the attributes whose calculated distance is the maximum possible from the respective *RP*;

• In dimensions without a chosen *RP*, the *AR* assumes a value 0 and does not affect analysis.

Finally, after processing the points, each of the records will have a value computed. This value is called *Data Relevance* (*DR*).

Definition 5: the *Data Relevance* (*DR*) is the computed value that describes how relevant a determined data item is, based on the *Attribute Relevancies* and on the *Max Relevance Distances*. For a given data item, the *DR* is the average of its correspondent *Attributes Relevancies*. For the *j*-th *k*-dimensional element of a data set, the DR_j is given by:

Where #R is the number of *Relevance Points*.

$$DR_j = \frac{\left(\sum_{i=0}^{k-1} PRC_j^i\right)}{\#R} \quad \textbf{(5)}$$

The *Data Relevance* value directly denotes the importance of its correspondent data element according to the user defined *Relevance Points*, and to visually explicit this fact, we use the *DRs* to determine the color and the size of the graphic elements. Hence, lower values stand for weaker saturations and smaller sizes, while the higher ones stand for more stressed saturations and bigger sizes, in contrast. Also, in our implementation we benefit from the fact that to denote relevance, only the saturation component is necessary, leaving the other components of the colors, hue and value, available to depict more information. Therefore, we projected a way to denote, along with the relevance analysis, the aforementioned frequency analysis.

That is, while the saturation of color and the size of the graphical elements denote relevance, the hue component of color presents the frequency analysis of the data set. More precisely, the highest frequencies are presented in red (hot) tones, and the lowest frequencies in blue (cold) tones, varying linearly passing through magenta.

Figure 6 presents the *Relevance Plot* over the Parallel Coordinates technique of the breast cancer dataset. In the three scenes we have defined 9 *Relevance Points* (dimensions 2 to 10). In 6(a) the points are set to the smallest values of each dimension, in 6(b) they are set to the maximum values of each dimension, and in 6(c) middle points are set.

In figure 6(a) the choice of the *Relevance Points* and its correspondent visualization leads to conclude that the lowest values of the dimensions' domains indicate class 0 (benign cancer) records. It also warns that this is not a final conclusion since the visualization reveals some records, in lower concentration, which are classified as 1 (malign cancer). It can be said that false negative cancer analysis is common with this clinical approach.

In figure 6(b) the opposite can be observed, the highest

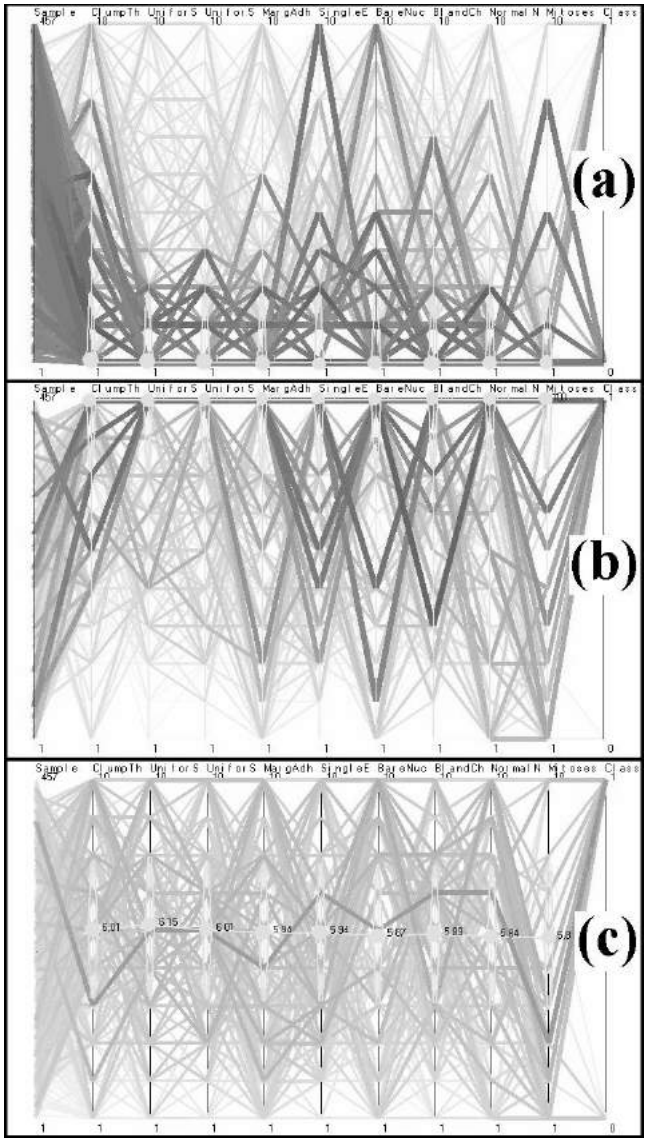

Figure 6 The Relevance Plot over a Parallel Coordinates scene. In (a) all the relevance points are set to the smallest values of their dimensions. In (b) they are set to the maximum values and in (c) middle points are set.

values indicate the records of class 1. It can be seen that false positive cancer analysis can occur, but they are very unusual, since just a shadow of pixels heads to class 0 in the 11th (right most) dimension.

Finally, in figure 6(c) the *Relevance Points* were set to middle points in order to make an intermediate analysis. Through the visualization, one can conclude that this kind of laboratory analysis is quite categorical, since just one record is positioned in the middle of the space determined by the dimensions' domains. But, in such cases, it is wise to classify the analysis as a malign cancer or, otherwise, to proceed with more exams.

7. Conclusions and future work

We believe that both the *Frequency* and the *Relevance Plot* might strongly contribute to improve the effectiveness of databases exploration. It is also expected that this contribution is applicable to other multivariate visualization techniques beyond the Parallel Coordinates and the Scatter Plots, specially the relevance visualization, that is simply a way to focus interesting parts of a data set without losing the overall sight.

Future work might be envisioned to extend the present one. The relevance visualization demands a certain processing power to be used interactively, and the described model does not embodies optimization nor scalability. Thus, research to develop visualization software that is scalable and able to efficiently hold other visualization issues are also expected.

Acknowledgments

The research presented in this paper was supported, in part, by the Sao Paulo State Research Foundation (FAPESP) under grants 01/11287-1, 02/07318-1, and by the Brazilian National Research Council (CNPq) under grants 52.1685/98-6, 860.068/00-7 and 35.0852/94-4.

References

1. Inselberg, A. and B. Dimsdale. *Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry*. in *IEEE Visualization*. 1990: IEEE Computer Press.
2. Bennett, k.P. and O.L. Mangasarian, *Robust linear programming discrimination of two linearly inseparable sets*, in *Optimization Methods and Software*. 1992, Gordon & Breach Science Publishers. p. 23-34.
3. Rundensteiner, A., et al. *Xmdv Tool: Visual Interactive Data Exploration and Trend Discovery of High Dimensional Data Sets*. in *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*. 2002. Madison, Wisconsin, USA: ACM Press.
4. Keim, D.A. and H.-P. Kriegel, *Visualization Techniques for Mining Large Databases: A Comparison*. IEEE Transactions in Knowledge and Data Engineering, 1996. **8**(6): p. 923-938.
5. Ward, M.O. *XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data*. in *Proceedings of IEEE Conference on Visualization*. 1994.
6. Fua, Y.-H., M.O. Ward, and A. Rundensteiner, *Hierarchical Parallel Coordinates for Exploration of Large Datasets*. Proc. IEEE Visualization'99, 1999.
7. Wong, P.C. and R.D. Bergeron. *Multiresolution multidimensional wavelet brushing*. in *Proceedings of IEEE Wsualization*. 1995. Los Alamitos, CA: IEEE Computer Society Press.
8. Keim, D.A., *Information Visualization and Visual Data Mining*. IEEE Transactions on Visualization and Computer Graphics, 2002. **8**(1): p. 1-8.
9. Bier, E.A., et al. *Toolglass and Magic Lenses: The See-Through Interface*. in *SIGGRAPH '93*. 1993.
10. Ahlberg, C. and B. Shneiderman. *Visual Information Seeking: Tight coupling of Dynamic Query Filters with Starfield Displays*. in *Proc. Human Factors in Computing Systems CHI '94*. 1994.
11. Keim, D.A. and H.-P. Kriegel, *VisDB: Database Exploration Using Multidimensional Visualization*. IEEE Computer Graphics and Applications, 1994. **14**(5): p. 16-19.
12. Card, S.K., J.D. Mackinlay, and B. Shneiderman, *Using Vision to Think*. Readings in Information Visualization. 1999, San Francisco, CA: Morgan Kaufmann Publishers.
13. Siirtola, H. *Direct Manipulation of Parallel Coordinates*. in *International Conference on Information Visualization*. 2000.
14. Kandogan, E. *Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions*. in *IEEE Symposium on Information Visualization 2000*. 2000. Salt Lake City, Utah.
15. Rao, R. and S.K. Card. *The Table Lens: Merging Graphical and Symbolic Representation in an Interactive Focus+Context Visualization for Tabular Information*. in *Proc. Human Factors in Computing Systems*. 1994.
16. Wegman, E.J. and Q. Luo, *High Dimensional Clustering Using Parallel Coordinates and the Grand Tour*. Computing Science and Statistics, 1997. **28**: p. 352–360.