

Data Exploration Tools for Multidimensional Data

Gabi Schmidberger
Department of Computer Science
University of Waikato
Hamilton 3240
New Zealand
gabi@cs.waikato.ac.nz

ABSTRACT

This paper presents new semi-graphical data exploration tools for multidimensional data. All new methods discussed herein are based upon histogram-like density estimation.

The new graphics are coarse visualizations of multidimensional histograms. The histograms were built using a new histogram method. This method splits the range into bins of varying lengths in such a way that the density in each bin is closest to univariate. In multidimensional space the algorithm finds bins that represent multidimensional areas of similar densities. This histogram method is used for simple representation tools which list the attributes of these bins and help the user to gain information about clusters and patterns in the data.

A further section drafts the design for a graphical user interface which displays pixel graphics and comprises interactive functions for data exploration.

Categories and Subject Descriptors

I.6.9.d [Visualization]: Multivariate Visualization; I.5.3 [Pattern Recognition]: Clustering; I.5.5 [Pattern Recognition]: Implementation—*Interactive Systems*

Keywords

Data Exploration, Data Mining, Density Estimation

1. INTRODUCTION

The first and obligatory step in data analysis is called ‘looking at the data’. This means for one-dimensional data to look at the values themselves or to generate summaries of the data like boxplot or histogram diagrams. Two-dimensional datasets are explored using a scatterplot and for three dimensional data a three-dimensional scatterplot can be generated. A popular diagram used for multidimensional data is the matrix plot, which is a matrix of scatterplots with all combinations of attributes on the axes. The more attributes a dataset has, the more difficult it is to find a representation

This paper was published in the proceedings of the New Zealand Computer Science Research Student Conference 2008. Copyright is held by the author/owner(s).

NZCSRSC 2008, April 2008, Christchurch, New Zealand.

of the data which helps uncover the existence of clusters and other characteristics of the distribution of the data.

With the increasing use of computers in all areas of life, more and more data is gathered and stored, which may be useful for later analysis. These datasets not only have several attributes but also thousands or millions of examples stored in their instances. The tools represented in this paper aim to support the exploration of large datasets with several thousand instances. In a scatterplot, if plotting the instances as black dots on white background, the user sees the local densities as the variation in the darkness of the area. The darkest areas are the densest. But if the number of instances is very high, some areas could be blacked out, and no information about density variation could be gained from the plot. The histogram obviates this problem of large datasets since it is a smoothing of the scatterplot. The density of the range that a bin stands for is the height of the bin, which is computed from the number of instances which fall into that bin n_i , the volume of the bin v_i and the total number of instances N .

$$D = n_i / (v_i * N)$$

Why use graphics for data analysis? Standard statistics forms an hypothesis and then uses statistical techniques on a data sample to test this hypothesis. Tukey and Tukey started with a new approach. They emphasise visualization of the data with the argument [2]: “There is nothing better than a picture for making you think of questions you had forgotten to ask.” The exploration of the data per graphic representation can make obvious which hypothesis should be put forward.

The visualization methods developed are based on a new technique of building histograms. For a traditional histogram the range of values is split into intervals of fixed length. The new method splits the range into intervals of varying length. The bin width is adapted to the underlying density distribution and a histogram which is built using this binning will show the most important features of the density distribution. In the multidimensional version that data is split parallel to the axes into multidimensional rectangles. The new simple graphic methods represented in this paper use this binning to support the user in exploring the data.

These bins are represented in a very simplified and coarse way, using just characters and not pixel graphics. The em-

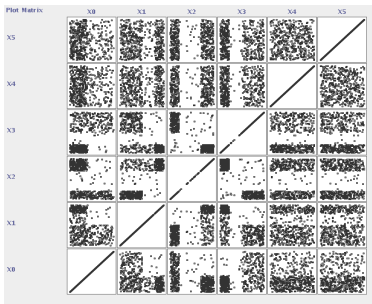


Figure 1: Matrix plot of the example dataset.

phasis is to give an overview of how the distribution of the data looks and to give an idea where a single bin is situated in the multidimensional range. A graphical user interface would make it possible to draw blocks of exact length. This would show more details but would also obscure the part of information the user needs to discover at this point of exploration. Of course further exploration steps would demand more detail. Future work could enhance this simple semi-graphic tool with a graphical user interface and so enable more detailed viewing of the data. Other further functionality could be interactive graphics.

My thesis project is the development of the new histogram method for one-dimensional and multidimensional data, and its application to density estimation and several machine learning techniques. The work on the data representation methods and pattern recognition methods described in this paper are not the main focus of my thesis but will be used to document my results in my written thesis.

2. TOOLS FOR DATA EXPLORATION

An overview of all existing graphic tools for data exploration of large datasets would go beyond the scope of this paper. Some well known methods are listed in Unwin, Theus and Hofmann [3]: Histograms, boxplots, scatterplots and parallel coordinates. The next sections will introduce two new semi-graphic representations for large datasets, the *semi-graphic bin list* and the *semi-graphic bin position overview*. The basis for these new graphics is a new binning method which was developed by Schmidberger and Frank (see [1]). It was used by them as a density estimator for one-dimensional data. In this work the method has been adapted for multidimensional data and the new method splits the data into bins of multidimensional rectangles building a multidimensional histogram.

The following sections explain the histogram method and introduce the two new graphics which are based on it. The last section adds the design of a fully graphical user interface which could support these diagram methods as tools for exploration.

2.1 The New Histogram Method

The new binning method splits the multidimensional range in a tree-like fashion. All splits are axis-parallel cuts. In a first step, the algorithm searches for a first split point by examining the range of each attribute to find the best split point. The decision regarding which is the best is based on a

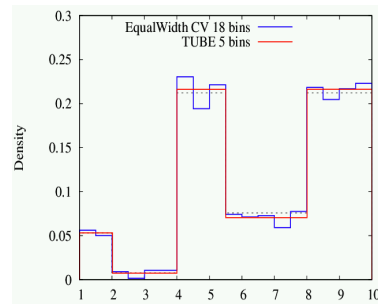


Figure 2: Varying bin width.

splitting criterion. The splitting criterion maximises at the point in each value range which reflects at best the change of density in the distribution of the instances. If the number of attributes is n , the algorithm then decides between these n split points again using a criterion and splits the range into two multidimensional bins. Schmidberger and Frank [1] explain the split criterion in more detail.

Next the algorithm repeats step one. It first suggests $2n$ split points then chooses one of them and splits one of the bins. These steps are continued until a fixed number of bins is produced. The number of bins is given by the user as a parameter. The result is a split tree. The root node in the tree is the first cut made, the nodes represent all further cuts and the bins are at the leaves of the tree.

Figure 2 shows the resulting histogram of a one-dimensional split compared to the histogram built with a conventional method which used a fixed bin width. The new histogram method adapts the bin width to the change of density. Like any histogram it can be used as an estimate of the real density function of this attribute.

2.2 Semi-graphic Bin List

The multidimensional histogram method splits the range into areas, choosing the sizes of the bins in such a way that they adapt to the local density. The resulting bins show the most significant features of the distribution of the instances. The *semi-graphic bin list* lists the bins with the most important features of the bin: Density, volume and the number of instances that the bin contains. The order of the bins is determined by gathering the leaves of the splitting tree from left to right (smaller subranges to larger value subranges). Note, in a multidimensional histogram the order of the bins is not clearly defined as it is for the one-dimensional histogram.

For the following examples a dataset was generated with 6 attributes and a few areas of uniform distribution (see Figure 1 for a matrix plot of this dataset). Figure 3 shows an example of a *bin list*. The representation of the bins is strongly simplified. Each bin is given a number for further reference. One line in the graphic lists the following values of one bin, Density (Dns), percentage of number of instances in the bin (Ins), and percentage of volume (Vol), forming three columns. All three values are represented in a semi-graphical way by using X characters. One X stands for 10 percent. The volume is given as percentage of the total volume, and the number of instances is the percentage of all

```
#0 : Dns:[      ] Ins:[      ] Vol:[X.....]
#1 : Dns:[X.....] Ins:[X.....] Vol:[XX.....]
#2 : Dns:[XXXX.....] Ins:[X.....] Vol:[X.....]
#3 : Dns:[XXXXXXXXXX] Ins:[XXXX.....] Vol:[X.....]
#4 : Dns:[X.....] Ins:[X.....] Vol:[X.....]
#5 : Dns:[X.....] Ins:[X.....] Vol:[X.....]
#6 : Dns:[XXXX.....] Ins:[XXXXX.....] Vol:[X.....]
#7 : Dns:[X.....] Ins:[X.....] Vol:[X.....]
#8 : Dns:[X.....] Ins:[X.....] Vol:[X.....]
#9 : Dns:[X.....] Ins:[X.....] Vol:[XXXXXXXX.....]
Percentage of instances presented(Dns): 100%
```

Figure 3: A sample bin list of 10 bins; Bin 0 is empty.

```
...
#15 : Dns:[<1E-3.....] Ins:[<0.01.....] Vol:[XXXXXXXX...]
Percentage of instances presented(Dns): 99.87%
```

Figure 4: Values below 0.1%.

instances. The percent values are rounded up to the next 10. So if the percentage is 53.0 percent, six X characters are drawn. Therefore in the example (Figure 3) the number of X characters seen for both values do not add up to ten as might be expected.

For the density values nothing like a total sum exists. For the presentation of the density the ‘highest’ bin is taken as 100 percent and the densities of the other bins as a percentage compared to this bin. Therefore the densest bin is represented with [XXXXXXXXXX], a string containing ten X characters.¹

All values are rounded up to the next 10, but values which are smaller than 0.1 are not shown as [X.], instead they are written as <0.1 or <0.01 continuing down to <1E-6. (See two examples in Figure 4.) In practise it is useful to have small values emphasised in the graphic.

In the last line the value after **Percentage of instances presented** refers to the density column and is the sum of all instances represented by all X characters. The presence of this value should help to detect a distortion of the histogram in case one of the bins was very narrow and with that its density value very high. In relation to this very dense bin the other bins get very low density values and become invisible, hiding the features of the dataset. The histogram methods will have to be rerun with some parameters like minimal bin width reset to avoid this distortion.

As a feature, the number of characters could be increased so that e.g. twenty X characters stand for 100% allowing the graphic to show more detail. Further features could be added with a fully graphical interface. Some ideas for future features are summarized in the later section ‘Future work’.

2.3 Bin Lists For Two Class Problems

Sometimes the instances of the dataset are classified with a nominal value. With an additional column the distribution of a binary class (two possible values i.e. ‘true’ and ‘false’) can be documented in the bin list.

The representation is again very coarse in order to give a quick overview. The middle ten characters give the density

¹Therefore always one bin in the density column must have a representation of [XXXXXXXXXX].

```
#5:
[XXXXXXXXXX][XXXXX...]
[XXX.....][...XX.....]
[XXXXXXXXXX][XXXXXXXXXX]

#6:
[XXXXXXXXXX][XXXXX...]
[XXX.....][.....XXXXX]
[XXXXXXXXXX][XXXXXXXXXX]
```

Figure 6: Positions of bin 5 and bin 6.

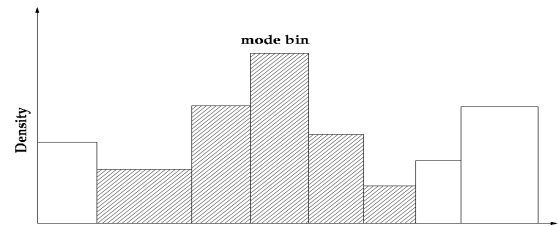


Figure 7: Cluster found in one-dimensional data.

of the first and second class ‘behind’ each other, with the class with the lower value to the front. The last character shows with uppercase A or B which of the two classes was denser in this bin. This is important if the two classes have the same number of characters in the middle part. The first character in the string shows which class was zero times represented in this bin, or is written 0 if neither was. When no instances at all are in this bin, the string in this column is [-.....-].

2.4 Semi-Graphic Bin Position Overview

The bin position graph is designed to give an overview of where in the total attribute range the bins are positioned. Figure 6 gives two examples of bin position overview graphs (for bin number 5 and for bin number 6 from the bin list example in Figure 4). For each attribute one character string shows the part of the range the bin covers. If the string is [XXXXXXXXXX] the bin was not cut in this attribute and covers its whole range. The string [XXXXX.....] means the range was cut in approximately the middle of the range and the bin covers the first part of the range.

In the given example the split tree never selected a cut in the last two attributes, so these attributes could be ignored for the presentation. The selection of attributes for presentation can be important if the number of attributes is very high.

2.5 Clustering

In the data exploration task it is often important to determine if and where clusters are in the data. Clusters are areas of high density surrounded by areas of lower density. The multidimensional binning is a good basis for clustering. It is an estimation of the density distribution of the data and therefore it is easy to determine the modes in the distribution. Self-evidently as a clustering algorithm a mode-seeking clustering algorithm was developed and implemented. The mode bin is a bin which is surrounded by bins with lower density. The algorithm finds cluster peaks and combines

```

#0 : -.....- Dns:[          ] Ins:[          ] Vol:[X.....]
#1 : Ab.....B Dns:[X.....] Ins:[X.....] Vol:[XX.....]
#2 : BaaaaaaaaA Dns:[XXXX.....] Ins:[X.....] Vol:[X.....]
#3 : ObbbbbbaaaaA Dns:[XXXXXXXXXXXX] Ins:[XXXX.....] Vol:[X.....]
#4 : Ab.....B Dns:[X.....] Ins:[X.....] Vol:[X.....]
#5 : Ab.....B Dns:[X.....] Ins:[X.....] Vol:[X.....]
#6 : Abbb.....B Dns:[XXXX.....] Ins:[XXXXX.....] Vol:[X.....]
#7 : Ab.....B Dns:[X.....] Ins:[X.....] Vol:[X.....]
#8 : Ab.....B Dns:[X.....] Ins:[X.....] Vol:[X.....]
#9 : Ab.....B Dns:[X.....] Ins:[X.....] Vol:[XXXXXXXX.....]
Percentage of instances presented(Dns): 100%

```

Figure 5: Bin list with information about class distribution.

them with its surrounding lower bins down to the ‘valleys’ of the distribution.

See Figure 7 for an example of a set of bins defined by the algorithm as a cluster (found in a one-dimensional dataset). To present the clusters in multidimensional data the user can either make *bin lists* of each cluster or use the *bin position graphic* with the function which combines all bins of a cluster to see the approximate position of the cluster in the data space.

2.6 Future Work

Both new graphic methods, the *bin list* and the *bin position overview* represent the data in a very coarse fashion. A future feature could be to give more precise values for the shown data in addition to the implemented version. To prepare for the design of these features, several scenarios of data exploration task should be developed. These could be exploring the areas of lowest or zero density in the data range or defining the areas with high density of positive instances and low negative instances and similar scenarios. The coarse graphics could give the user a first overview. A mouse click on one of the bins could open a new window with the more precise information. Another way would be to give optional view settings that could easily be changed, maybe like a scale setting in a text file that changes the zoom of the text.

Several graphics should be combined with each other. If the bin is selected in one graph it is also highlighted in the other. This should also be combined with a matrix plot where the ranges of the selected bin get drawn into the scatter plots. Figure 8 is only part of a matrix plot of a dataset with more than 100 attributes. It is obvious that the information, where the positive instances in this dataset are the densest can not be seen easily. With the new tool the rectangle of this area could be drawn into the matrix plot.

A hierarchical nesting of the binning could be enabled. A bin can be selected and the histogram algorithm could define a histogram with a given number of bins using the instances in that bin only. The clustering could look for density modes within the bin.

The aim is to combine the already implemented graphics with flexible functionality and a fully graphic user interface to build an effective tool for the support of the data exploration task.

3. CONCLUSIONS

These very simple graphics represent the multidimensional histogram and give a coarse overview of the structure of the

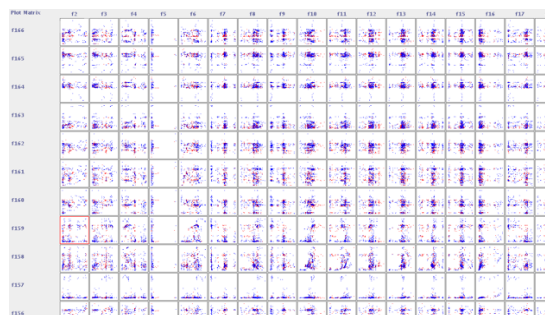


Figure 8: Part of a matrix plot of a dataset with 166 attributes.

data. The graphics can be quickly generated. The simplicity of the information guides the user to the main features of the data. The diagram methods are not restricted by the number of instances in the data and can also be applied to very large datasets.

The described methods for representing the result of the binning will be used to document my thesis work. In my thesis I developed the binning method and applied it as density estimator to several machine learning algorithms. So far I have used the new visualization methods only for documentation purposes. The deployment of the graphics in data exploration tasks would be interesting to see, but probably would require the implementation of a well thought-out toolset which utilizes a graphical user interface.

4. REFERENCES

- [1] G. Schmidberger and E. Frank. Unsupervised discretization using tree-based density estimation. In *Proc 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 240–251, 2005.
- [2] J. Tukey and P. Tukey. Computer graphics and exploratory data analysis: An introduction. In *Sixth Annual Conference and Exposition: Computer Graphics*, pages 773–785, 1985.
- [3] A. Unwin, M. Theus, and H. Hofmann. *Graphics of Large Datasets*. Springer, New York, 2006.