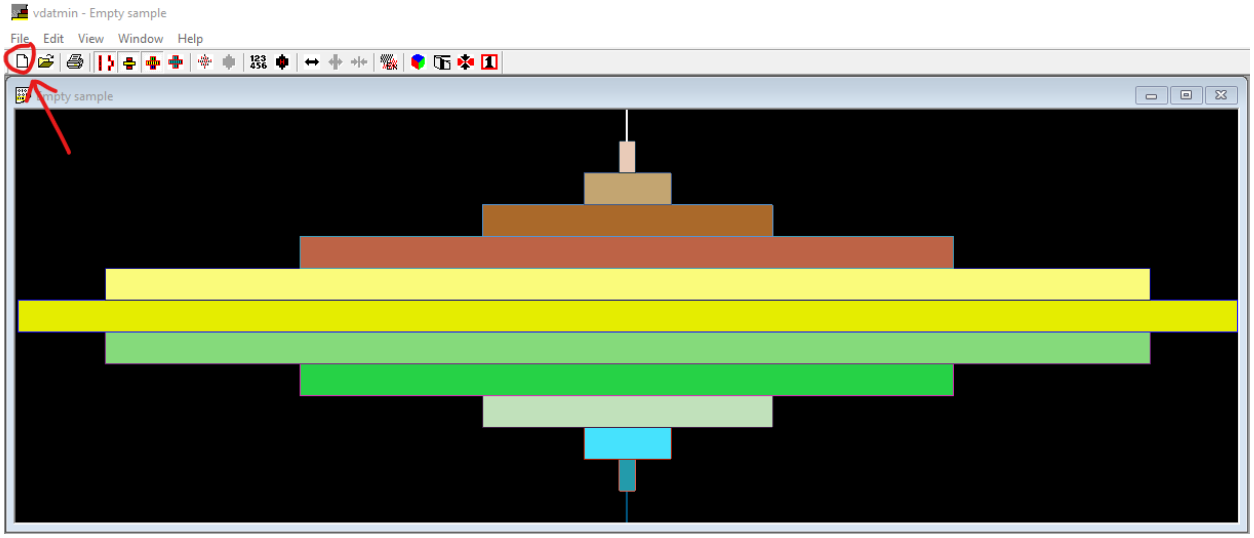
Introduction

Vdatmin is an application that plots binary vectors onto a series of disks, which is called the Multiple Disk Form (MDF). The MDF form without any vectors can be seen when first opening up the application. Each disk represents the Hamming norm of the vector, which is the sum of each attribute of the vector. Therefore, the number of disks is the dimension of the vectors plus one. A particular vector is plotted on the disk that corresponds to its Hamming norm. The default position of a vector is determined by its decimal value. Vectors with large decimal values will be plotted on the right side of a disk. The bottommost disk accounts for a Hamming norm of 0, whereas the highest disk accounts for the maximum Hamming norm (every attribute of a vector is 1). For example, given a vector of (1101), the Hamming norm is 3, and the vector will be plotted on the rightmost side of the 4th disk out of a total of 5 disks. Vectors are plotted onto a disk as vertical bars. Vectors with a class of 0 are white bars, and vectors with a class of 1 are black bars.

Step 1

Run vdatmin.exe, then click on the "New" button (marked with a red box and red arrow).

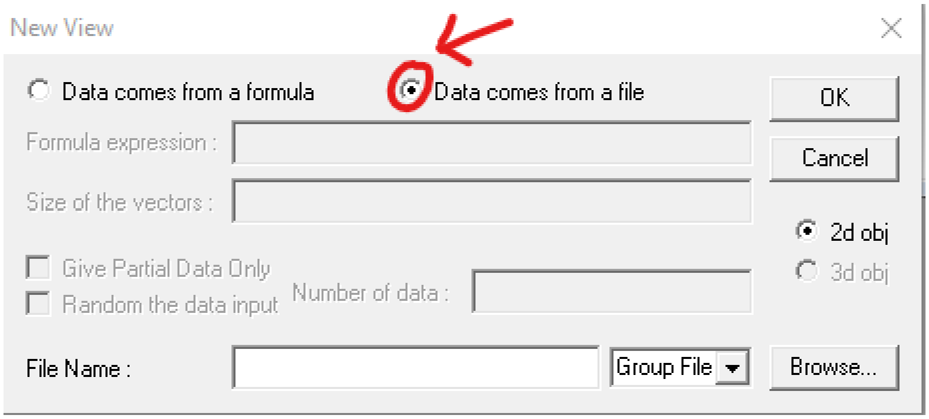
<u>Note:</u> in the "File" tab, there are several standard Microsoft Office features such as "Open": ignore these.



Step 2

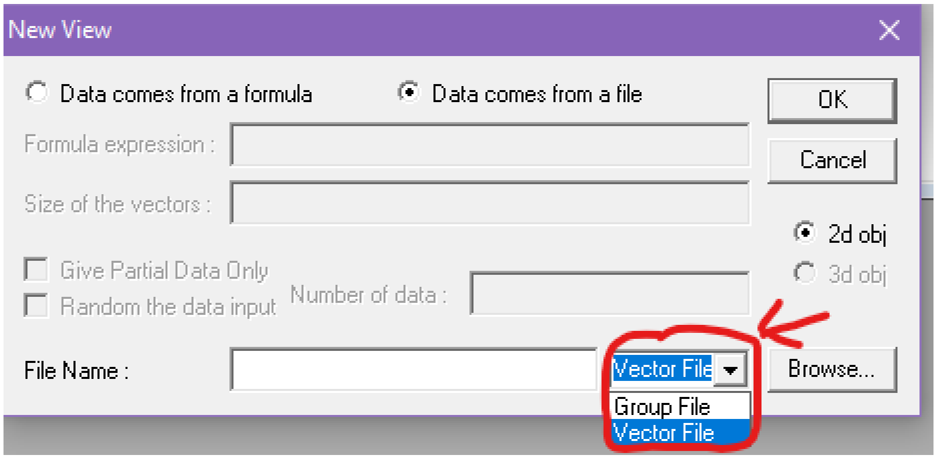
Click on the “Data comes from a file” toggle.

<u>Note:</u> the “Data comes from a formula” feature is experimental and will crash the program.



Step 3

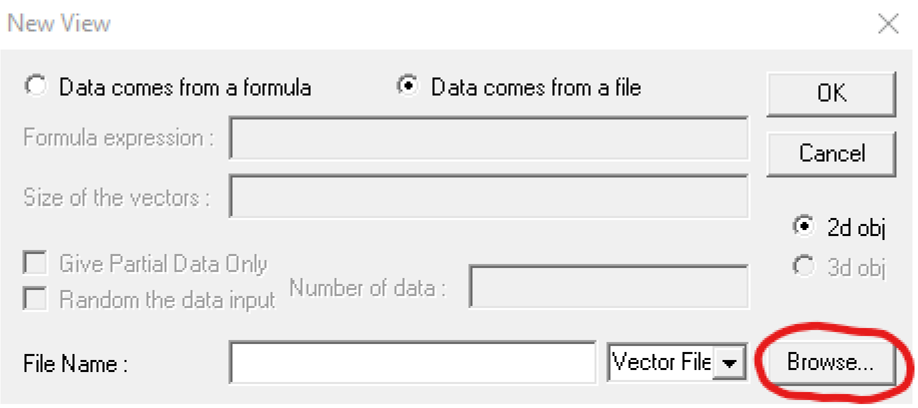
In the drop-down menu, select “Vector File.”



Step 4

Click on “Browse” to browse for the “cancer10.bvect” file in the “/datasets/” directory, which is located in the folder that this application is in. Then, click on “Open” once locating the file.

<u>Note:</u> to view the vectors in this dataset, change the file extension to “.txt” and open the file in Notepad or a similar text editor. You could edit the dataset to see the differences in visualizations after completing this tutorial. Make sure to not accidentally add newline or invisible characters to the end of the file, and also make sure to change the file type back to “.bvect” by saving with the “All files” option enabled in the drop-down for file types.



New View

○ Data comes from a formula ● Data comes from a file

Formula expression :

Size of the vectors :

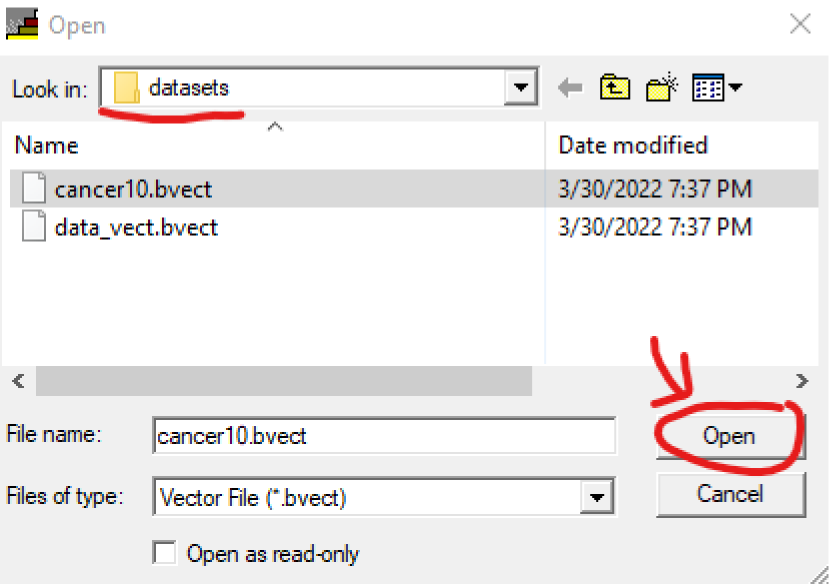
● 2d obj ○ 3d obj

□ Give Partial Data Only

□ Random the data input

Number of data :

File Name : Vector File



Open

Look in: datasets

Name	Date modified
cancer10.bvect	3/30/2022 7:37 PM
data_vect.bvect	3/30/2022 7:37 PM

File name: cancer10.bvect

Files of type: Vector File (*.bvect)

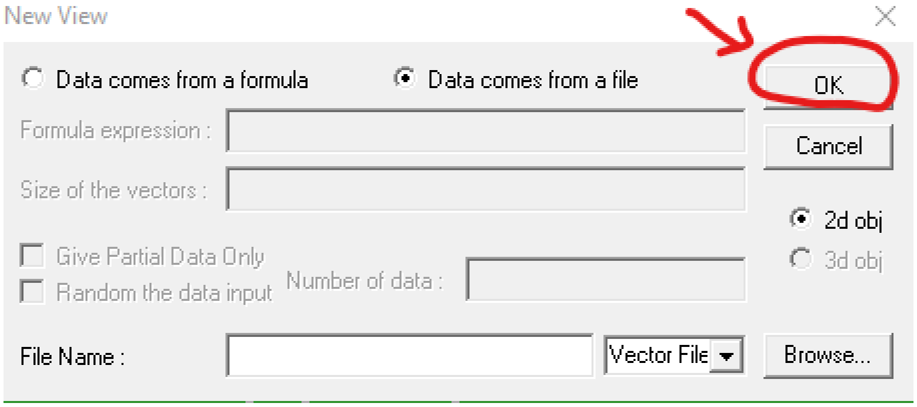
□ Open as read-only

Definition 4.1 – cancer dataset

This dataset contains vectors with 11 attributes that represent benign tumors (non-cancer) or malignant tumors (cancer). The last attribute is the class of the vector. Each attribute is a true or false statement about an attribute of a case of a tumor. Therefore, vectors with a class of 1 (black bars) are malignant tumors, and vectors with a class of 0 (white bars) are benign tumors.

Step 5

Click on "OK" to visualize the vectors.



New View

○ Data comes from a formula ● Data comes from a file

Formula expression :

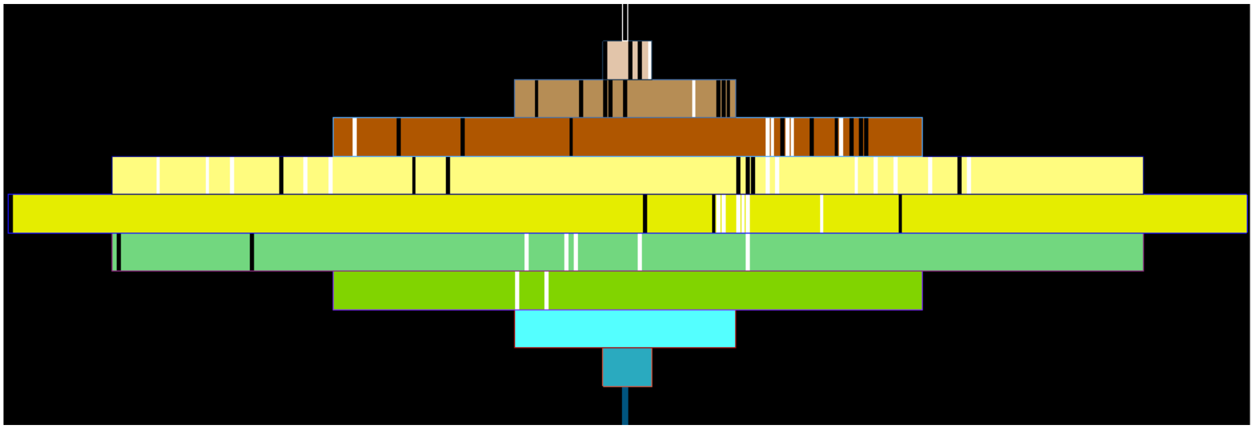
Size of the vectors :

● 2d obj ○ 3d obj

□ Give Partial Data Only

□ Random the data input Number of data :

File Name : Vector File

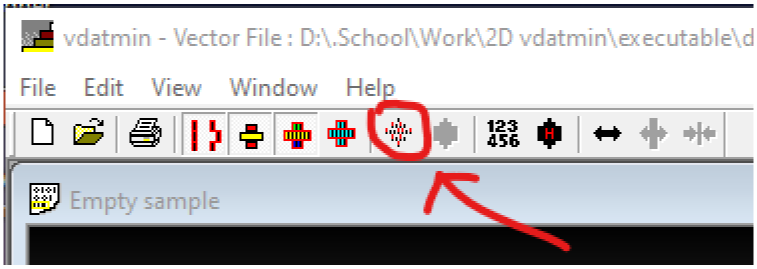


Visualization Observations: notice the separation of data. We can see that black bars (cancer) seem to trend towards the upper disks, but that is not always the case. Therefore, we can hypothesize that vectors with more attributes of 1 (true) are more likely to represent cancerous tumors. It also seems like there are more white bars on the right sides of the disks.

Step 6

Now we will use some of the tools to change the visualization.

Click on the “Bush-up” button.



vdatmin - Vector File : D:\.School\Work\2D vdatmin\executable\d
File Edit View Window Help
123
456
Empty sample

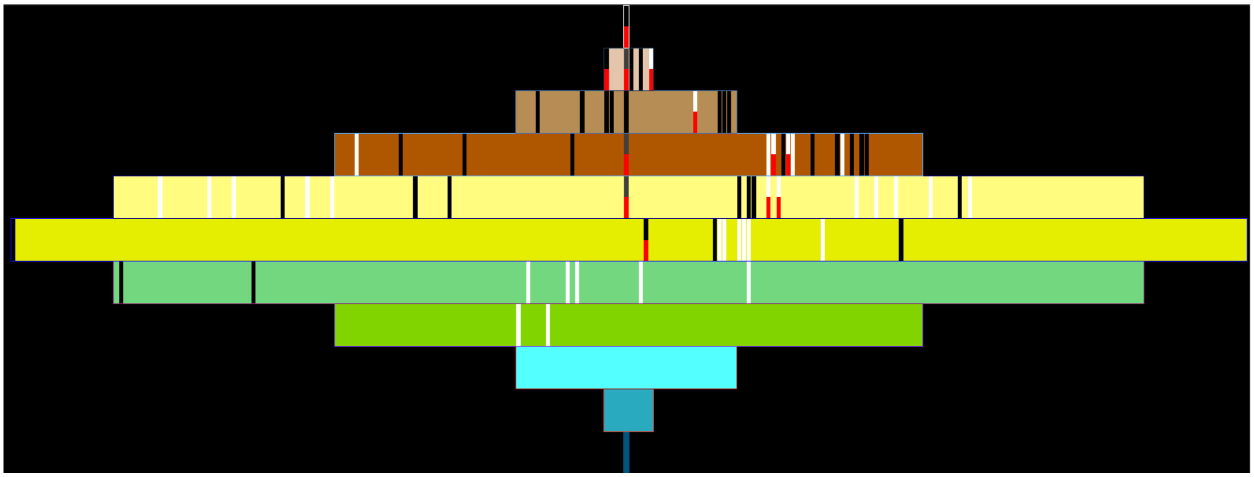
Definition 6.1 – bush-up

A bush-up of a vector consists of all vectors that are greater than the selected vector. Therefore, the bush-up of a vector will always consist of vectors on disks that are greater than the selected vector’s disk. However, not every vector on a greater disk will be selected.

Step 7

Click on a vector (bar) to show the bush-up of that vector. The below picture indicates what the visualization should look like by selecting the middlemost vector of the “cancer10.bvect” file.

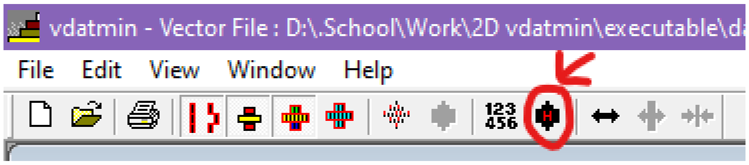
__Warning:__ after selecting a vector, please avoid clicking on a vector with the same horizontal position to avoid possibly crashing the program. For most vectors, this is not a problem.

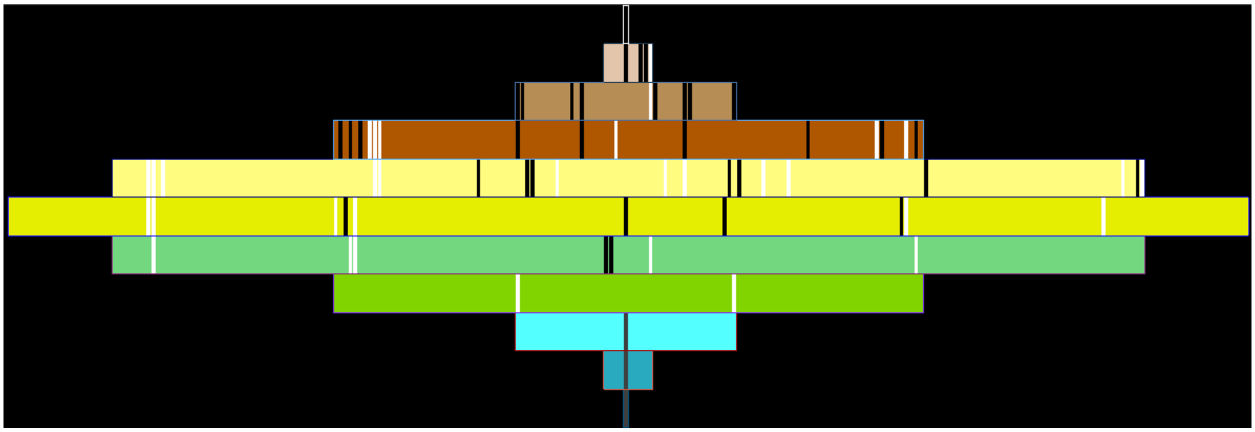


Step 8

Next, we will visualize this dataset with Hansel Chains.

Click on the "Hansel Chain" button. To switch back to the decimal order visualization, click on the button directly to the left of the "Hansel Chain" button.


vdatmin - Vector File : D:\.School\Work\2D vdatmin\executable\d
File
Edit
View
Window
Help
123
456



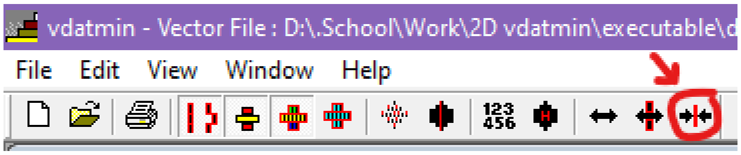
<u>Visualization Observations:</u> in addition to our previous observation of black bars (cancer) tending towards being on the upper disks, it also seems that they are slightly more closely grouped than white bars (benign).

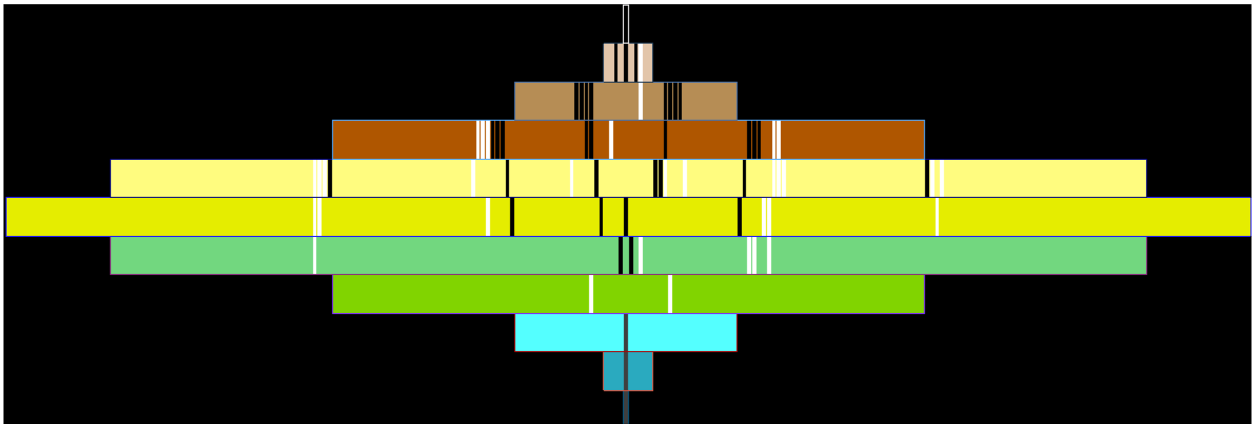
Definition 8.1 – Hansel Chains

Remember that the default position of a vector in MDF form is calculated with the decimal value of a vector. Instead, Hansel Chains can be used to calculate the position of the vector. Hansel Chains refers to a set of Hansel Chains. A Hansel Chain is a sequence of vectors, where each vector is greater than the preceding vector, but only one attribute differs between consecutive vectors. For example, (000) < (001) < (011) < (111) is a Hansel Chain. Every vector in a dataset belongs to some Hansel Chain. This idea can be explored in detail in Dr. Boris Kovalerchuk's paper, "**<u>Consistent and complete data and 'expert' mining in medicine</u>**," in addition to the attached paper, "Visual Data Mining and Discovery with Binarized Vectors."

Step 9

To center the Hansel Chains to get a clearer visualization, click on the "Center Hansel Chains" button.

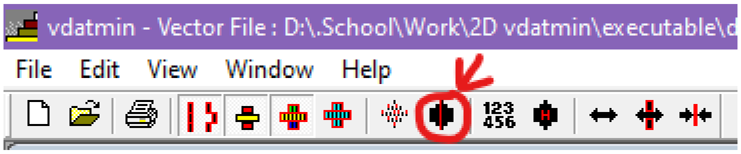

vdatmin - Vector File : D:\.School\Work\2D vdatmin\executable\d
File
Edit
View
Window
Help
123
456

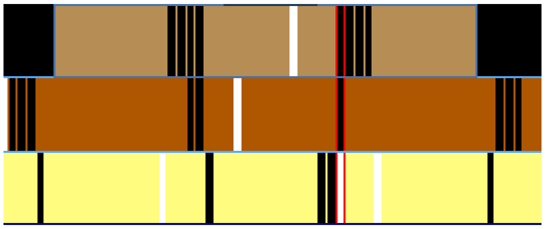


<u>Visualization Observations:</u> centering the Hansel Chains seem to confirm our previous observation: that black bars (cancer) are more tightly grouped. White bars tend towards being on the outer Hansel Chains. The next step will make it easier to see how what vectors are in what Hansel Chain.

Step 10

To highlight a specific chain that a particular vector is in, click on the "Show Hansel Chain" button, then select a vector (roughly the center of a bar must be clicked). A particular chain will be highlighted in red.

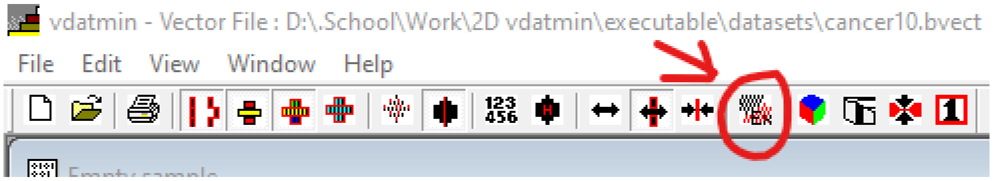

vdatmin - Vector File : D:\.School\Work\2D vdatmin\executable\d
File
Edit
View
Window
Help
123
456

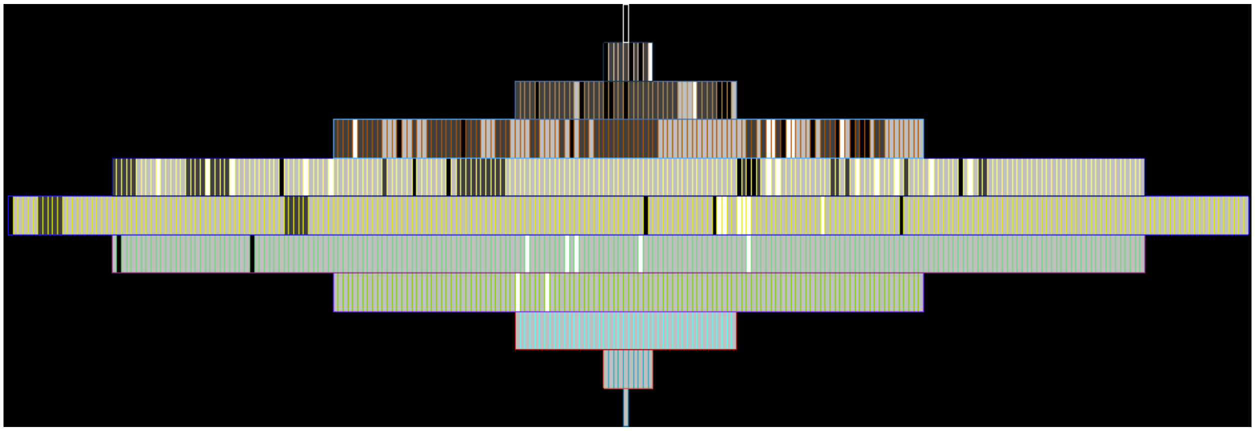


Step 11

Next, we can show every possible vector for the dataset by clicking on the "Monotone Expansion" button.

<u>Note:</u> once this feature is toggled, it cannot be toggled off without reading the file over again.


vdatmin - Vector File : D:\.School\Work\2D vdatmin\executable\datasets\cancer10.bvect
File
Edit
View
Window
Help
123
456
Empty sample

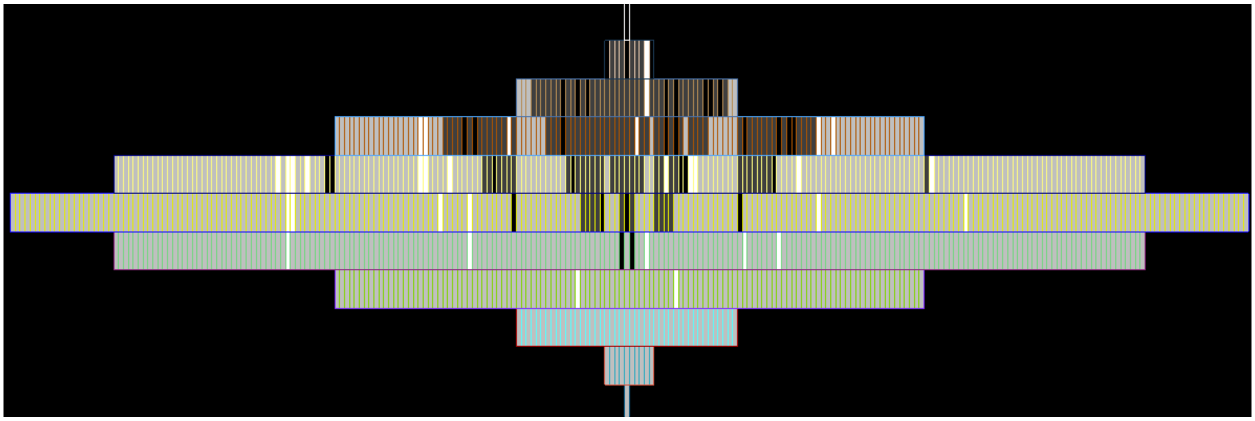


Definition 11.1 – Monotone Boolean Expansion

With this visualization, we also see light gray and dark gray bars that take up the entire MDF. These bars are expanded vectors, meaning that a vector ***y*** can be derived from a vector ***x*** as long as the monotonicity hypothesis holds. If a vector ***y*** is less than a vector ***x*** and ***x*** has a class of 0 (white), then ***y*** is a light grey bar. If a vector ***y*** is greater than a vector ***x*** and ***x*** has a class of 1 (black), then ***y*** is a dark grey bar. These expanded vectors account for every possible vector for a given dimension for the vectors. To read more about monotone Boolean expansion, refer to Dr. Boris Kovalerchuk's paper, "**<u>Consistent and complete data and 'expert' mining in medicine,</u>**" or the paper attached in this directory, "Visual Data Mining and Discovery with Binarized Vectors."

Step 12

Next, let’s use what we learned to sort these expanded vectors with Hansel Chains. Then, center the Hansel Chains.



<u>Visualization Observations:</u> it seems that our previous observations were correct. Remember that dark grey bars represent expanded vectors with a class of 1 (cancer), and that light grey bars represent expanded vectors with a class of 0 (benign). Here, it is much easier to see that cancerous tumors tend to have more attributes that form a particular pattern. If we find a cancer patient with attributes that matches one of these black or grey bars, then if the monotonicity property holds, we can conclude that the patient has cancer (also assuming that the dataset is correct).

Step 13

This tutorial is over! There are several other buttons, such as moving a selected vector or plotting in 3D instead of 2D. Use these with your own discretion. When hovering over a button, the purpose of the button is given in the bottom left of the application window. To move the view of a 3D visualization, “WASD” rotates the disk upwards, leftwards, downwards, and rightwards. “Z” zooms outwards, and “X” zooms inwards. “Q” and “E” shift the visualization to the left or right.