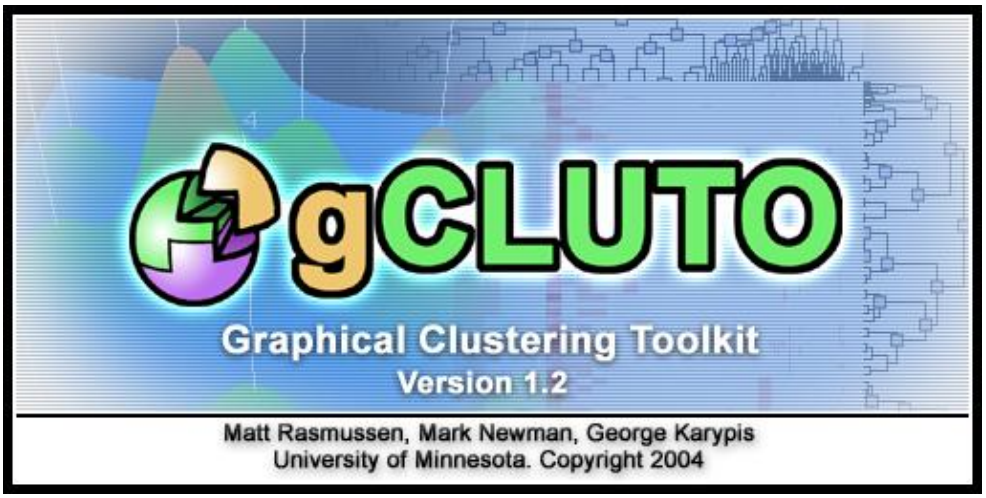

gCLUTO Documentation

Version 1.2

Matt Rasmussen, Mark Newman, George Karypis
University of Minnesota. Copyright 2004
rasm0146@umn.edu, karypis@cs.umn.edu
October 7, 2004

Table of Contents

1 Introduction **1**
- 1.1 What is gCLUTO 1
- 1.2 Features 1
- 1.3 Version Information 1
  - gCLUTO 1.2 − October 7, 2004 1
  - gCLUTO 1.1 − June 20, 2004 1
  - gCLUTO 1.0 − November 11, 2003 1
  - gCLUTO 0.5 − January 20, 2003 2

2 Installing gCLUTO **3**

3 Using gCLUTO **4**
- 3.1 Overview 4
- 3.2 Creating a New Project 4
- 3.3 Importing Data 5
- 3.4 Clustering Data 6
- 3.4.1 Comparing Solutions 7
- 3.5 Visualizing Solutions 7
  - 3.5.1 Matrix Visualization 7
  - 3.5.2 Mountain Visualization 10
- 3.6 Printing 11
  - 3.6.1 Printing Details 11
- 3.7 Exporting 12

1 Introduction

1.1 What is gCLUTO

gCLUTO (Graphical CLUstering TOolkit) is a graphical front−end for the CLUTO data clustering library. Its purpose is to make CLUTO's clustering abilities available in a user−friendly and graphical way. In addition, gCLUTO provides several ways to interactively visualize clustered results. A copy of gCLUTO can be found at http://www.cs.umn.edu/~mrasmus/gcluto. For more information about CLUTO visit http://www.cs.umn.edu/~karypis/cluto.

1.2 Features

gCLUTO has the following features:

- Project tree that manages data files, clustering solutions, and visualizations.
- Detailed dialogs for choosing clustering options
- Spreadsheet interface for viewing data
- HTML interface for viewing solutions.
- Bootstrap clustering
- Matrix Visualization − a colored interactive matrix
- Mountain Visualization − a 3D visualization generated using Multidimensional Scaling.
- Printing and Exporting data and visualizations

1.3 Version Information

gCLUTO 1.2 − October 7, 2004

- Added exporting of solutions to Partition Vectors
- Added importing of partitioning from external clustering algorithms
- Fixed reveresed color coding in Mountain Visualization
- Fixed dense graph importing

gCLUTO 1.1 − June 20, 2004

- Dialogs updated and stored in XML resource file
- Projects can open from command−line
- Reduced disk space requirements for projects
- Object/Feature search in DataView and MatrixVis
- Data Properties Dialog
- Bootstrap Clustering

gCLUTO 1.0 − November 11, 2003

gCLUTO has been extended and reorganized to make it a more productive tool. The following features have

been added to gCLUTO 1.0:

- Reorganized project directories
- Files are deleted when corresponding items are deleted from project
- Long operations performed in a background thread
- Importing tab, space, etc. delimited files into projects
- Exporting data/solution matrix to tab delimited files
- Exporting Solution Report to HTML file
- Solution columns in Data View
- Sorting in Data View
- External Clustering Quality Statistics in Solution Reports
- Right click information in Matrix Visualization
- Matrix Visualization labels stretch with available space
- "View All Objects" and "View Only Clusters" added to Matrix Visualization
- Mini−Solution Report in Mountain Visualization
- Printing Matrix and Mountain Visualizations to printers and files

gCLUTO 0.5 − January 20, 2003

gCLUTO is currently in an alpha phase. The purpose of this release is to explore what features and user interface designs would work best for a clustering application.

2 Installing gCLUTO

Currently, gCLUTO is available for both Linux and Microsoft Windows platforms. To install gCLUTO:

- Find the latest version of gCLUTO at http://www.cs.umn.edu/~mrasmus/gcluto.
- Download and unzip the archive to any location on your computer.
- Read the *README.txt* file to locate the correct version of gCLUTO for your Operating System.
- Windows users can make a desktop shortcut to *gcluto.exe* by locating gcluto.exe in the file manager, right−clicking the icon, dragging it to the desktop, and choosing "Create Shortcut Here" from the pop−up menu.
- Linux users can create a symbolic link to the gcluto binary ("ln −s gcluto wherever/you/want/the/link"). Place the symbolic link wherever is most convenient (ex: ~/bin).

Note: the actual executables (gcluto, gcluto.exe) must stay within their folders in order to work properly. Do not relocate them. gCLUTO locates its supporting files (icons, images, etc.) by directory paths relative to the executable's location. Relocation of the executable can cause these files to not load properly.

---

3 Using gCLUTO

3.1 Overview

When clustering data, many pieces of information are involved, such as data files, clustering solution files, and visualizations. Like many other applications, gCLUTO uses the concept of a **project** to organize the user's data and work flow. When a project has been loaded, its contents will be displayed in the tree view located at (a) in Figure 3.1.

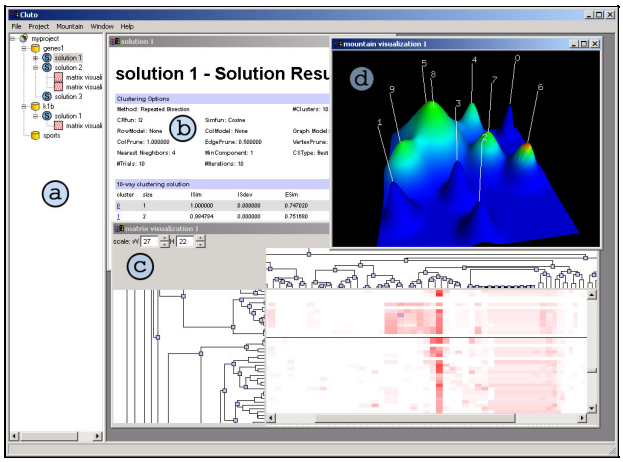

Figure 3.1 − screen shot of gCLUTO

Each item in the project is presented as an icon in the tree.

- **Project** − This represents the project itself. It is the root of the project tree.
- **Data** − After importing data into a project, one of these icons will appear in the project tree. A project can contain many different data items.
- **Solution** − After clustering one of the data items, a solution item will be created and placed underneath the data item from which it was clustered.
- **Matrix Visualization** − This is a visualization that can be generated after clustering. All visualizations appear under the solutions from which they were generated.
- **Mountain Visualization** − This is another visualization that attempts to describe the interrelationships of clusters in a 3D way.

Right clicking on any item will bring up a pop−up menu listing the available operations that can be preformed on the item. **Double clicking** on any item will open its contents in a new window called a **view**, similar to the windows (b), (c), and (d) in Figure 3.1. While working in one of these views, extra menu options specific to the view's content will appear in the menu bar.

3.2 Creating a New Project

When gCLUTO first opens it starts with an empty project tree. To begin work, a new project must be created. To create a new project, go to the menu bar and choose "File" and then "New Project". A file dialog window will appear. Specify a name for your project and a location on your computer to save it.

gCLUTO will create a directory, called the **project directory**. The project directory will be named after the project and stored at the specified location. Within the project directory, gCLUTO will save all the information related to the project.

To open an existing project, choose the "File" menu and then "Open Project". A file dialog will appear. Navigate to the location of the project directory and open it. Within the project directory there will be a file named "*project_name*.prj", where *project_name* will be the name of the project. Choose this file and click "Open".

After these steps, a project will be loaded and displayed in the project tree.

Note: gCLUTO 1.0 uses a different project directory structure than gCLUTO 0.5. Projects created using gCLUTO 0.5 can not be opened in gCLUTO 1.0 and vice versa.

3.3 Importing Data

As of gCLUTO 1.0, gCLUTO accepts three file formats: CLUTO matrix files (*.mat), CLUTO graph files (*.graph), and dense matrix delimited files. For details on the exact formats of CLUTO matrix and graph files see CLUTO's documentation.

The following file types are used when importing data in CLUTO file formats:

- **matrix file (*.mat)** − contains a dense matrix, sparse matrix, or similarity graph representing the data to be clustered.
- **row labels (*.rlabel)** − contains labels for the rows of the data matrix.
- **column labels (*.clabel)** − contains labels for the columns of the data matrix.
- **class labels (*.rclass)** − contains class labels for the rows of the data matrix.

Delimited files can be created by hand or exported by most spreadsheet programs. gCLUTO can accept tab, space, semicolon, and comma delimited files. Other characters can also be specified as delimiters.

To import a new data item go to the menu bar and choose "Project" and then "Import Data". The **Import Data** dialog will appear allowing the user to specify the location of a file for each of the file types listed above. Clicking on a "Browse" button will bring up a file dialog to allow the user to locate the needed files. Only the *.mat file is required. The user must also specify whether the *.mat file contains matrix data or graph data by selecting the appropriate option.

If the *.mat file is chosen first, gCLUTO will try to guess the location of the optional files (*.rlabel, *.clabel, *.rclass) by appending the appropriate extension onto the *.mat filename. For example, for a file named *genes.mat*, gCLUTO will guess *genes.mat.rlabel* for a row label file. If such a file exists, gCLUTO will make it the default file to open in the "Browse" file dialog. Using this file naming convention can simplify the importing process.

If the user chooses to import a delimited file, the delimited file options will become enabled. gCLUTO can optionally interpret the first row in the delimited file as column labels. In addition, gCLUTO can optionally interpret the first column as row labels. User may also specify which characters should be used as delimiters. If multiple characters are specified, then the occurrence of any one of them will cause a field delimitation. Blank fields (two delimiters with no data in between) are allowed in delimited files. If a blank occurs where a number is expected, then it will be interpreted as a zero. If a blank occurs where a label is expected, then a

default label of "no−label" will be used.

After specifying these files, the user may give a label for the data item. If no label is given, the data item will be labeled after its *.mat file with the extension removed. After clicking "OK" in the Import Data dialog, gCLUTO will attempt to read in the chosen files. If no errors are encountered, gCLUTO will add the new data item to the project tree and open a **Data View**. The Data View allows the user to view the data and verify that it has been loaded correctly.

3.4 Clustering Data

If data has been imported using the steps given in 3.3 then it is ready to be clustered. Clustering can be initiated two different ways. The first is choosing "Cluster" from the pop−up menu that appears when you right−click on a data item in the project tree. Secondly, the very same menu can be found in the menu bar under "Data" if a Data View is open.

After choosing "Cluster" in either menu, a **Clustering Options** dialog will appear with all the options available for clustering. These options work exactly the same as in CLUTO. For an explanation of their meanings see CLUTO's documentation. Only particular options make sense together. To help make sensible choices, gCLUTO will automatically update the dialog as the user makes choices to ensure that only reasonable choices are available.

gCLUTO can now import clustering solutions derived from external sources. To apply such a clustering to a dataset in gCLUTO, choose the "External" clustering method in the Clustering Options Dialog and click "Cluster". This will prompt a Open File Dialog where the user can specify a Patition Vector file. A Partition Vector file is an ASCII text file that specifies a partitioning. For an explanation of the file format see solution export format in section 3.7.

Once the clustering options are chosen, click "Cluster" in the Clustering Options dialog. After gCLUTO finishes the clustering calculations, it will respond by creating a solution item under the clustered data item in the project tree.

gCLUTO will also automatically open a **Solution View** similar to (b) in Figure 3.1. This view contains the options used for clustering and several statistics about the clusters. The report is designed after the report given by CLUTO. For further explanation of its meaning see CLUTO's documentation. In addition, the report contains links, similar to a web page. Clicking on these links allows for quick navigation between related information in large reports.

gCLUTO has been designed to facilitate clustering of the same data multiple times. If a previously clustered data item is chosen for clustering again, the Clustering Options dialog will appear with the options that were used the previous time. To reload the options used for creating a particular solution, right−click the desired solution item in the project tree and choose "Recluster" from the pop−up menu. This will bring up the Clustering Options dialog with the solution's options loaded. This feature eases the process of repeated adjustments to clustering options.

3.4.1 Comparing Solutions

The Data View can be used to compare multiple solutions for a single data item. When solutions exist, a second spread sheet, called the **Solution Grid**, will appear on the right hand side of the view. Each column of the Solution Grid belongs to a different solution and is labeled after the solution's label shown in the project tree. The integers appearing in the Solution Grid signify to which cluster each row of the data matrix was partitioned.

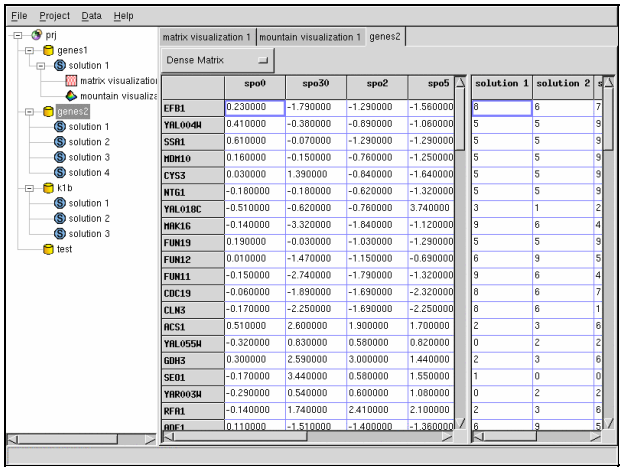

Figure 3.2 − screen shot of the Data View

The Solution Grid and the **Data Grid** to its left are designed to work together. Thus, scrolling vertically in either grid will cause the other to also scroll. Row resizing is also synchronized between the two grids. Horizontal scrolling is done independently in order to allow comparisons to be made between any column in the Data Grid and any column in the Solution Grid.

Clicking on any column label in the Data View will sort the rows of the Data Grid and Solution Grid based on the values of the column. Clicking on the same column twice will reverse the direction of the sort. Therefore, if a column label in the Solution Grid is clicked, rows of the same cluster will be displayed together. This allows quick comparisons of the partitioning of different solutions. Sorting can be reset by choosing the "Reset Sorting" option in the "Data" menu. Columns in the Data Grid cannot be sorted when the grid is in **sparse matrix** view.

3.5 Visualizing Solutions

Currently, gCLUTO contains two visualizations: the **Matrix Visualization** and the **Mountain Visualization**. Visualizations can be generated from solutions by choosing the desired visualization from the solution menu. This menu can be found by right−clicking on a solution item in the project tree or in the menu bar under "Solution" if the user is currently working in a Solution View.

3.5.1 Matrix Visualization

The Matrix Visualization is similar to the matrix visualization produced by CLUTO. The former extends the latter by making the matrix interactive. A detailed explanation of the visualization is given in CLUTO's documentation.

In the Matrix Visualization, the orginal data matrix is displayed such that colors are used to graphically represent the values present in the matrix. gCLUTO uses white to represent values near zero, increaingly darker shades of red to represent large values, and increasingly darker shades of green to represent negitive values. The rows of the matrix are reorder, such that rows of the same cluster are together. Black horizontal dividers separate the clusters.

If tree building is enabled, the Matrix Visualization will contain trees located above and to the left of the matrix. If an agglomerative clustering algorithm was used, the tree generated during clustering is displayed as the **Row Tree**. Otherwise, a tree is generated to fit the clustering solution. The **Column Tree** is generated by performing agglomerative clustering on the inverse of the matrix.

NOTE: Tree building of large datasets can have large run−time and memory requirements. If the Matrix Visualization generation is taking too long, try disabling tree generation.

If row and column labels were chosen when the data was imported, then they will appear below and to the right of the matrix. Labels will only show if space is available to display them.

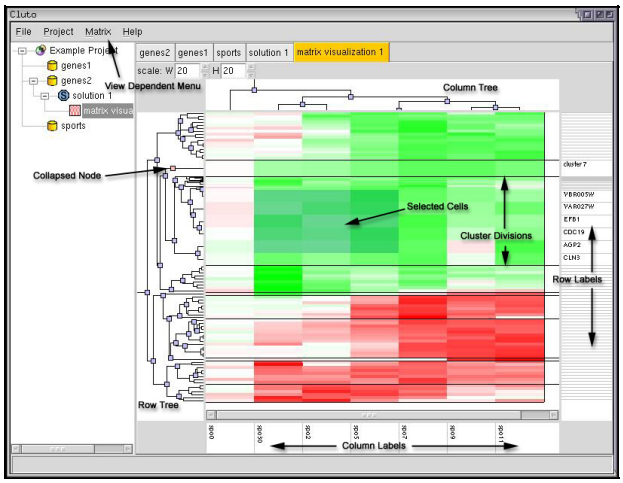

Figure 3.3 − screen shot of the Matrix Visualization

To help explore the information contained within the Matrix Visualization, several features have been implemented. First, the size of the matrix can be scaled in multiple ways. Second, the trees can be used to collapse and expand areas of interest within the matrix. Lastly, the user can access information about any cell by clicking the right mouse button on cells. A pop−up dialog box will appear giving the row and column labels, the data value, and cluster id associated with the cell.

3.5.1.1 Matrix Visualization − Scaling

The easiest way to scale the matrix is with the scaling controls located directly above the matrix. Scaling can be changed by entering a new size in the text box, or by clicking on either of the up or down arrows. The control labeled with "W" controls the width of the matrix and the control labeled "H" affects the height. These scaling controls change the dimensions of the entire matrix and are convenient for zooming in and out of areas of interest in the matrix.

Often times the user needs to enlarge one area of the matrix, yet shrink areas that are not as important. This type of scaling can also be done. To resize only a portion of the matrix, start by selecting the area to be resized. Selection is performed by clicking on any cell and draging the mouse to another cell. These two cells will become the corners of the selected rectangular region. Cells that are selected are shaded blue. To resize the selected region, place the mouse over any edge of the region. The cursor will change to a resizing cursor. Click and drag the edge to the desired location. The selected cells will then resize to fit within the new region.

The following options are available under the "Matrix" menu:

- **Reset Sizing** − Restores all scaling to the default values
- **Fit to Screen** − Resizes the matrix to fit within the visualization window

3.5.1.2 Matrix Visualization − Using the Trees

The Row and Column Trees allow for collapsing and expanding of the matrix. Blue squares in the tree represent nodes that are fully expanded. Clicking on any expanded node will collapse it. Collapsed nodes are represented as pink squares. When a node is collapsed, all of its descendents are hidden. If a node in the Row

Tree is collapsed, all of the rows of the collapsed region are hidden and replaced with a single row that contains their average. Simply click a collapsed node to expand it again. The Column Tree works in a similar manner.

The labels will change to describe the collapsed regions. If a region contains rows which all belong to the same cluster, then it will be labeled with the cluster id. If multiple clusters are present in a collapsed region then it will be labeled "multi−cluster".

The following options are available under the "Matrix" menu only when the row tree has been generated:

- **View All Objects** − Expand all nodes in the row tree so that every object is visible
- **View Only Clusters** − Collapse nodes in the row tree so that only cluster averages are displayed

The "View Only Clusters" option is convenient for large datasets that are difficult to view when every row is displayed. Cluster averages also help depict the general trends present in each cluster.

3.5.2 Mountain Visualization

The **Mountain Visualization** is used to visualize the relative similarity of clusters as well as their size, internal similarity, and internal deviation. In the mountain visualization, each cluster is represented as a peak in the 3D terrain. A peak's location, volume, height, and color are all used to protray information about the associated cluster.

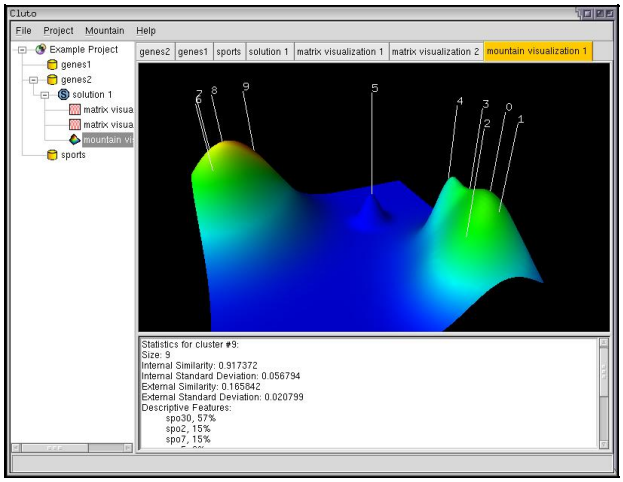

Figure 3.4 − screen shot of the Mountain Visualization

The user can navigate through and around the 3D visualization by clicking and dragging the mouse over the 3D display. Different mouse buttons perform different actions.

- **Left Click** − Rotates the terrain.
- **Right Click** − Moves the terrain up, down, left, and right.
- **Middle Click** − Zooms in and out.

The location of the peaks in the plane is determined using **Multidimensional Scaling** (MDS) on each of the cluster mid−points. MDS attempts to preserve the distances between vertices as they are mapped from a high dimensional space down to a lower dimensional space. In this application, cluster mid−points are used as vertices in MDS and are mapped to a two dimensional plane.

MDS allows users to make inferences about their data using the Mountain Visualization. For example, in Figure 3.4 a data matrix was clustered into ten clusters. The Moutain Visualization represents these ten clusters as ten peaks labeled by their cluster id. Although ten clusters were requested, MDS has placed the peaks in two distinct groups. We can infer that clusters within each group are strongly similar, while widely different from clusters in the other group. Thus, the visualization suggests the data would better lend itself to a two−way clustering.

The shape of each peak is a Gaussian curve. This shape is used as a rough estimate of the distribution of the data within each cluster. The height of each peak is portional to the cluster's internal similarity. The volume of a peak is portional to the number of elements contained within the cluster. The resulting Gaussian curves are added togther to form the terrain of the Mountain Visualization.

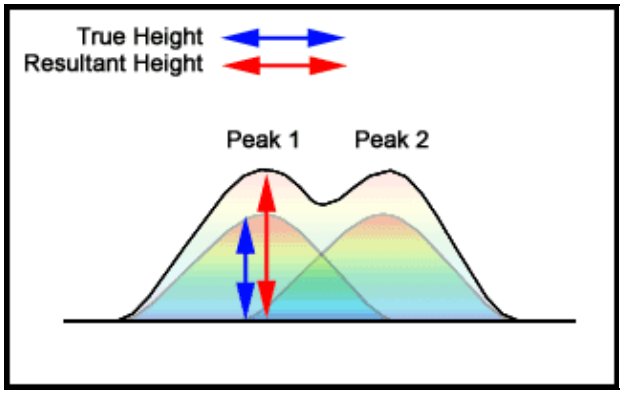

Figure 3.5 − How peak heights add

Note: When comparing peak heights keep in mind that the Mountain Visualization has added the peak curves together. As seen in Figure 3.5, the resultant height is

taller than the true height.

The color of a peak is proportional to the cluster's internal deviation. Red indicates low deviation where as blue indicates high deviation. Only the color at the tip of a peak is significant. At all other areas, the color is determined by blending to create a smooth transition.

Clicking on any label will load statistics about the associated cluster into the text window located below the visualization. This information is identical to the information found in the Solution Report. If column labels have been chosen for this data, then the Mountain Visualization can display the most common features above each peak. This option is called "Show Features" and is found in the "Mountain" menu.

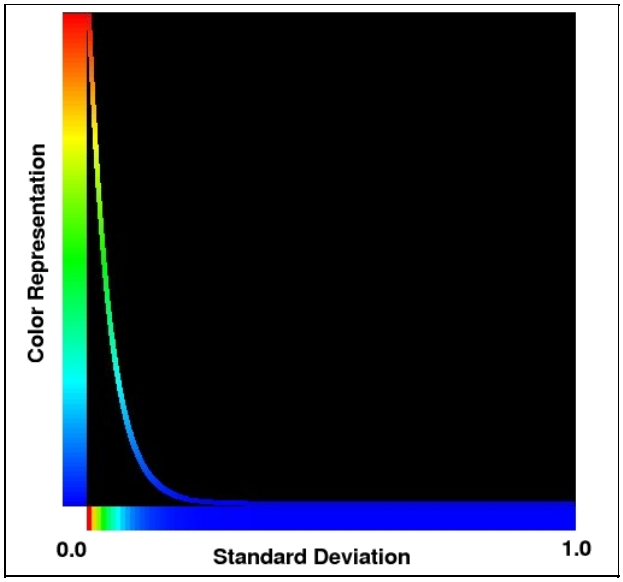

Figure 3.6 − Color code for standard deviation of clusters

3.6 Printing

Printing is performed in gCLUTO by first opening a View to print. Once a View is opened, printing and previewing can be initiated under the "File" menu. Printing is only available for the following Views:

- Matrix Visualization View
- Mountain Visualization View

Additional printing options such as page margins, page dimensions, and printer settings can be found by choosing the "Page Setup" and "Print Setup" options under the "File" menu. Views can also be printed to Post Script files by selecting the "Print to File" option in the Print Dialog.

3.6.1 Printing Details

The various dimensions of the visualizations printed are calculated by the dimensions of the **Printable Area**. The Printable Area is the area of the page that is within the margins. Choosing a smaller printable area will generate smaller graphics.

When printing the Matrix Visualization, the matrix is scaled to fit the entire Printable Area. Relative scaling of rows and columns is left intact. If labels or trees are present, they are each allocated roughly one sixth of the width of the Printable Area.

Mountain Visualizations are printed using the perspective and dimensions of the current Mountain Visualization View. The visualization is scaled to fill as much of the Printable Area as possible without exceeding the margins.

3.7 Exporting

Exporting is performed in gCLUTO by first opening a View to export. Once a View is opened, exporting can be initiated by choosing the "Export" option under the "Project" menu. The following Views can be exported:

- Data View
- Solution View (Solution Report)

Data Views are exported as tab delimited files containing the data matrix as well as any solution columns that are present. Row and column labels are also included in the first column and row, respectively. If labels are not given, then default labels "rowX" and "colX" are used, where "X" is replaced by the row or column number. Tab delimited files can be opened by most spreadsheet programs.

Solution Views can be exported as HTML files or Partition Vectors stored in ASCII text format. The HTML files will contain the same information that is present in the Solution Report. This makes it easier to use and share the information given in Solution Reports. Web browsers and most word processors can open HTML files.

The Partition Vector format allows gCLUTO's clusterings to be easily used in further analysis outside the software package. The format is an ASCII text file with a signle column of integers in the range 0 to the number of clusters minus one. The integer on the ith line of the parition vector specifies the cluster id of ith object in the dataset. Partition vectors can also be imported by gCLUTO in the cluster options dialog.

Matt Rasmussen rasm0146@umn.edu
Last Modified: Sat Oct 9 10:50:35 CDT 2004