Tumor Data Visualization

Anudeep Nallamothu
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
nan4@umbc.edu

Kishan Pawar
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
kish5@umbc.edu

Abstract:

Immunotherapy works better for some types of cancer then other types of cancer. One of the major problems with immunotherapy treatment process for cancer is finding a pattern out of available historical data and comparing which therapy is giving best results. Data Visualization helps to represent and study data for patterns, clusters and comparisons. We are proposed the visualization to show the comparison between the temporal data and the multivariate data. The visualization was created by the discussion with the faculty and the end users. The components used in the final visualization are bar charts and radial effect in the circular fashion. The most important feature about the proposed visualization is the comparison facility of the temporal data.

Keywords: temporal, multivariate, parallel co-ordinates, temporal dimension

1. Introduction:

Immunotherapy is the widely used treatment process for the cancer that uses certain parts of the person's immune system to fight disease like cancer. There are many types of therapies available. It is a major research question to identify which of type of therapy is effective treatment for what kind of cancer. To do this task, data visualization can be used. Creating visualization can be used to

So we wanted to create visualization with temporal dimension. We want study the effective therapies and any patterns in the data.

2. Related Work:

2.1 Temporal Data:

Temporal data denotes the evolution of object characteristics over a period of time. It is simply a data that represents a state in time, such as the land-use patterns of London in 2012, or total rainfall in New York on August 21, 2009. Temporal data is collected for the various applications and accordingly to take the decisions. For example, temporal data is taken to analyze the weather patterns and other environmental variables, monitor traffic conditions, study demographic trends, analyze the patient's health condition during the recovery state from a disease, and so on. This data comes from various sources from manual data entry to data collected using observational sensors or generated from simulation models. Temporal datasets are ubiquitous but notoriously hard to visualize, especially rich datasets that involve more than one dimension in addition to time. There are so many different methods todays that it has become hard for both researchers and designers to get a clear picture of what has been done, and how much of the design space of temporal data visualization remains to be

explored. In our work we are addressing the questions of the temporal data visualization problem for the cancer tumors as 1. Which treatment was effective at a certain time? 2. Which tumor responded well for which treatment at given time? We will be working on the tumor dataset which has the tumor types, tumor masses and the treatments. Figure 1 shows the example of visualization of the temporal data for the crime rates in the USA in 2012.

2.2 Multivariate:

Multivariate data analysis refers to any statistical technique used to analyze data that arises from more than one variable. This essentially models reality where each situation, product, or decision involves more than a single variable. The information edge

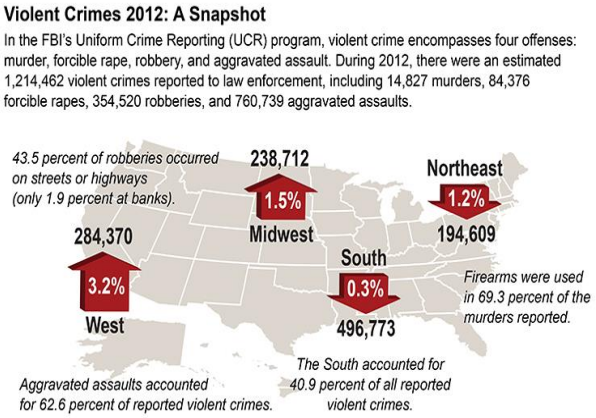


Figure 1. 2012 Violent Crimes in USA region wise.

has resulted in masses of data in every field. Despite of the amount of data available, the ability to obtain a clear picture of what is going on and make intelligent decisions is a challenge. Multivariate analysis can be used to process the information in a meaningful manner. We are using the multivariate analysis to determine the relation between the two variables in terms of the tumors and their respective treatments in the tumor dataset. The multivariate analysis of the given data and the visualization of the same is challenging as we have to deal with the multiple variables at the same time and have to study the change in the behavior of the variables from one state to other. Figure 2 shows the multivariate data visualization for all flights going to London. The different geographic areas are shown by different colors. The big concentric rings indicate the average daily number of flights from Milan to all worldwide cities which are directly connected. The area of each circle is proportional to the airport traffic of the city(cities with more than one airport report the total traffic). On similar basis we are going to implement the multivariate data visualiza-

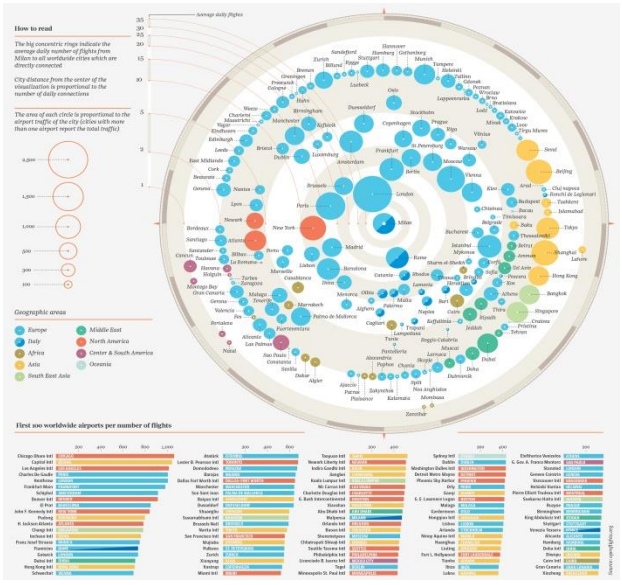


Figure 2. Multivariate data visualization flights connecting to London.

-tion based on the different values of the tumors masses and the treatments associated with them.

2.3 Parallel Co-ordinates:

Parallel co-ordinates is a visualization technique used to plot individual data elements across many dimensions. Each of the dimensions corresponds to a vertical axis and each data element is displayed as a series of connected points along the dimensions or axes. They are very powerful tool for understanding multidimensional numeric datasets. Figure 3 shows the parallel co-ordinates visualization for the car models released from 1970 to 1982 and contains their mileage (MPG), number of cylinders, horsepower, weight, and the year they were introduced. The parallel co-ordinates will have its utility in our work to visualize the tumor mass values and the treatments associated with individual tumors.

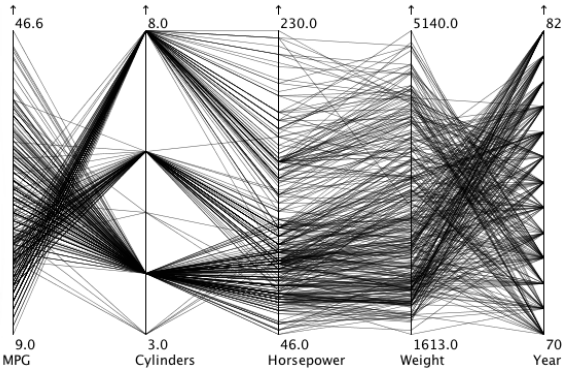


Figure 3. Parallel co-ordinates visualization for the car models released from 1970 to 1982.

3. Design Methods:

3.1 Data Abstraction:

Data used in the tumor data visualization is the dataset which consists of the following aspects: 9 treatments, 18 different protein types and the 2 nodes. The data is in the text and numbers format where numbers are the floating point numbers/values of the tumor masses and the values of those tumor masses with respect to the above mentioned treatments. The data is clean in nature and has more quantitative nature.

Since the data is quantitative in nature, ant data pre-processing like data cleaning was easy and it was easy to interpret as it did not have any missing values. The tumor masses indicated the effective values to that particular treatment from the given treatments in the dataset.

3.2 Task Abstraction:

The initial phase of the visualization was to analyze and identify the tasks associated with the dataset provided. The task analysis was challenging in terms of the deciding the appropriate visualization for the particular tasks. The major task in the initial phase was to find the effective treatment amongst the given treatment for a particular protein or tumor. This task was challenging to visualize, since the final visualization should be able to show the end-user or the doctors/practitioners to find the effective treatment as mentioned above. This actual process for the visualization for this task began with the discussion with the faculty. In initial phase the first visualization we came up with, had some drawbacks in terms of the user interactivity and the ease of understanding. Then the analysis and detailed discussion lead us to finalizing the visualization which had minimum drawbacks in terms of the visual aspects. Thus finding the effective treatment is our first task. This task will lead in answering the questions such as what treatment is effective on a particular tumor and which is not.

The second major task of our visualization was to find the temporal relationship between the different treatments with proteins. The

dataset we had consisted the data of the two years viz., 2015 and 2016. Thus the major challenge was to visualize the visualization that will be able to show the temporal comparison between the treatments and the proteins. The aim of this task was to show the trend of how one particular treatment is effective or not on a particular protein over a period of time. The temporal comparison between two factors was our second major task of the visualization.

3.3 Design Consideration:

We are proposing the multicomb visualization, spike glyph and scattered plot for visualizing the temporal, parallel co-ordinates and multivariate data.

3.3.1 Multicomb visualization:

Multicomb visualization is the circle shaped visualization with the axes in its circumference. Each axis signifies the importance of the data variable in the given dataset and its values. Accordingly on each axis the data points are plotted and they are joined with the vectors. Thus the complete visualization as a whole looks like a multiple combs is places in a circular fashion. Each comb/axis in our work will show the treatment and its values of the corresponding tumors. The axis has two sections, on each section we are planning to visualize the 2015 and 2016 year's data values present in the given dataset and to compare them with each other. This will show us the effective treatment for a specific tumor. Figure 4 shows the proposed visualization.

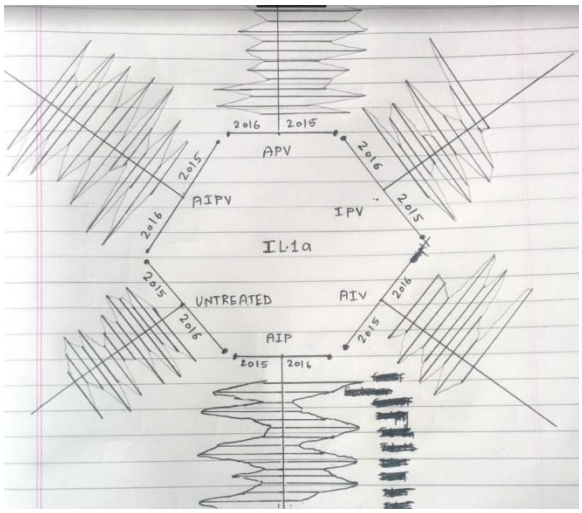


Figure 4. Multicomb visualization

3.3.2 Spike glyph:

The spike glyph is same as multicomb visualization, but the only difference is that on each axis instead of bidirectional plots, it has the regular bar chart plot. This bar chart plot looks like the spikes all around the circle. On each of the axis we are going to plot the treatment and its corresponding tumor values on the bar chart. Figure 5 shows the proposed visualization. The spike glyph looks like the spikes all over the visualization and it is widely used in the mining industries according to their requirements.

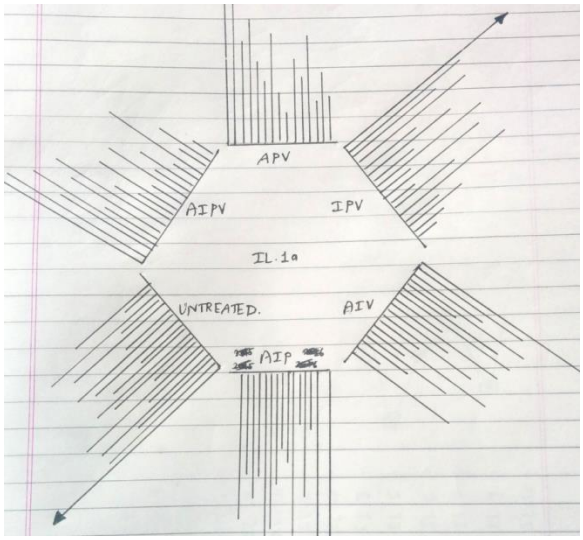


Figure 5. Spike Glyph

3.3.3 Scattered plot:

Scattered plots are similar to line graphs in that they use horizontal and vertical axes to plot the data points. However, they have a specific purpose. Scattered plots show how much one variable is affected by another. Thus because of this property of the scattered plots, we have chosen this visualization to show the multivariate data. The relationship between the two variables is called their correlation. Figure 6 shows the proposed visualization. The scattered plot visualization will be an interactive visualization where user can click on each of the data points to see the actual data points for the corresponding variables. The active data point will be nothing but the average of all tumor masses for corresponding treatment. Thus on user click on that average value will let him/her see the actual values for the tumor masses for the corresponding treatment. The vertical axis will have the treatment and horizontal axis will have the different tumor masses.

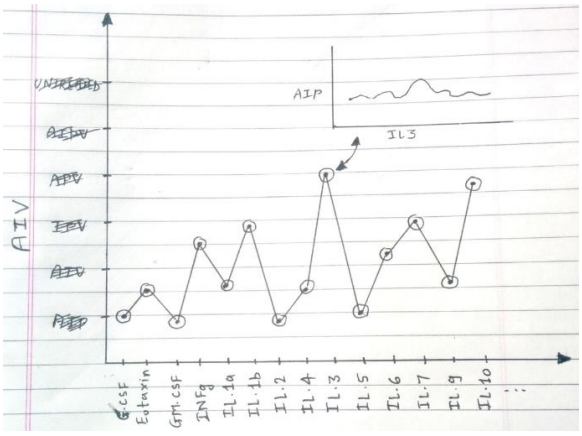


Figure 6. Scattered plot visualization

As the proposed designs were implemented and the further research was carried out, we finally came up with the one new design. The new design was proposed with the collaboration with the end users perspective and the guidance of the faculty. The new design was the combination of the multicomb visualization and the spike glyph. The basic design motivation behind the design was the bar charts and its power.

Now-a-days the bar charts are one of the very powerful visualization itself and it gives more visual detail if combined with other visualizations. Bar charts is a graph that presents grouped data with the rectangular bars with lengths proportional to the values that they represent. The bars can be plotted horizontal or vertical. Bar charts have the discrete range. They are usually scaled so that all the data can fit on the chart. Thus focusing on this property, we stacked the 18 different proteins in a way that they can be compared according to the users need treatment wise.

The new proposed design also contains the radial effect since it is in the circular manner and has many grids to compare the data with. The design was proposed considering the temporal data that we have and the designs we were working on. The radial effect is a graphical method of displaying multivariate data in the form of two dimension charts of three or more quantitative variables represented on axes starting from the same point. The relative position and the angle is typically uninformative. New design has the circular structure same as the multicomb visualization with the radial effect.

The data we had consisted of the 18 different proteins and 9 different treatments according to the tumor masses. Thus we decided to break down the 18 different protein values in the less number of clutters so that it will be an ease to understand the visualization.

This decision was made because of the temporal data and the stacked bar charts property. The temporal data is shown as stacked bar charts and side by side and thus it created a lot of visual clutter and the confusion analyzing the visualization. Thus the new proposed idea dealt with this problem and visualization was easy to understand.

The changes in the design were constant as development proceeded, but we tried to stick to the initial proposed design and not deviate from that. Since we were implementing the radial effect to the visualization, we constantly faced difficulties while implementing it. The challenges were the coding challenges and the design limitations. The radial effect visualization has the limitation of making trade-off decisions as when one chart is greater than another on some variables, but less on others. Further, it is hard to visually compare lengths of different combs, because radial distances are hard to judge though concentric circles help as grid lines according to grids. Instead, one may use a simple line graph, particularly for time series.

The cluttered grids of the proteins were arranged as per the temporal data that we had. The temporal data was year wise which was 2015 and 2016 year data. All the design challenges were the implementation challenges and the on time modification changes in the visualizations by discussing with the faculty.

4. Implementation Method:

The tumor data visualization was implemented using following:

4.1 Display:

1920*1080 resolution display.

4.2 Input Devices:

The input devices for the visualization was the QWERTY keyboard and and the track pad.

4.3 Hardware:

Intel i7 x64 bit with 8GB DDR3 RAM and 4GB Graphics Card.

4.4 Software:

The following software was used:

- Sublime Text
- Python 3.4

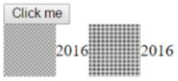
4.5 Language:

The following languages were used:

- Data Driven Document(D3)
- Python
- CSS
- HTML
- JavaScript

5. Results and analysis of results:

The final visualization was as shown in figure 7 below:



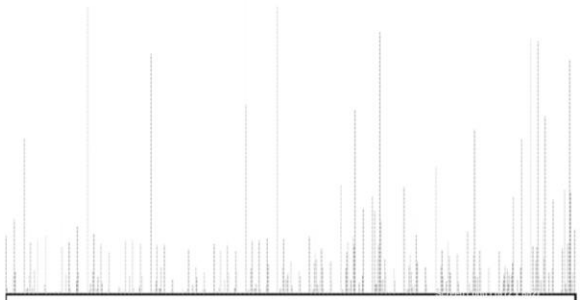


Figure 7. Final Visualization

Figure 8. Radial Shape Visualization

The final result was the visualization with addressing the tasks mentioned in section 3.2. The final visualization showed the temporal comparison and was able to interactively show the effective treatment among the given treatment for a particular tumor or protein. The final visualization consisted of the radial shaped bar chart so called multicomb visualization as mentioned above. Temporal data comparison feature for particular year. Trend to analyze the effectiveness of a treatment on a particular protein.

The proposed design had the grids as levels on the radial shape for the different proteins, but down the timeline later in development phase the discussion with faculty made us to remove the grids. Thus the final visualization don't have the grids on the radial shape for different proteins while the stacked bar charts were shown on the same line horizontally across the radial shape. The initial design also had the bar charts filled with the solid colors, but in final design the bar charts were filled with the textures as discussed with the faculty and according to the design requirements. In final visualization the dial effect is not present as there were lots of problems in implementing the dialer effect in terms of rotating the entire design with the different grid levels.

6. Discussion of results:

The final visualization analysis was in terms of the following aspects:

- No grids were visible as levels on the radial shape for different proties.
- Was able to compare the temporal data
- Was able to find the effective treatment.
- Was able to show the trend with the time.

7. Challenges:

The challenges were the implementation challenges as mentioned above and the design changes while going further. The design changes in the later phase of the development played the significant role as we had to constantly make changes in the codes and test them again and again. Thus the major challenge was the coding challenge.

8. Future Work:

The future work for this visualization is to implement the dialer to compare different proteins in the different grids by providing the user interactive interface to rotate the different grids. The rotation should be on the button click event in one both directions. So incase user want to compare the three protein values with other three particular protein values then he should be able to select those six values and compare them by rotating them using the dialer feature of the visualization. The other major advancement for this particular visualization would be the user interactivity for the selection of the different aspects of the design such as colors, angle of the rotating dialer and so on.

9. Conclusion:

Thus with the help of the visualization the doctors or the practitioners will get to know about the effective treatment. After the complete implementation we'll be able to show the relationships between the different attributes of the tumor data. Because of the temporal data visualization, will be able to show the nature of the treatment or tumor at a particular instance of time. The visualization will be helpful to take the necessary measures on the improvement of the treatment on a particular tumor.

10. References:

[1]Visualization of multivariate data.(http://people.stat.sc.edu/hansont/stat730/MultivariateDataVisualization.pdf)

[2]https://circlebloom.com/cancer-visualization-complete-program/

[3]http://www.practicalradonc.org/article/S1879-8500(13)00066-0/abstract

[4]https://books.google.com/books?id=gFnsQlXmQBsC&pg=PA56&lpg=PA56&dq=immunotherapy+visualization&source=bl&ots=tzVfWzMAdK&sig=G70KQ2F3QnApeUXfFIhR8gtu0Ow&hl=en&sa=X&ved=0ahUKEwiy14eCgpjQAhUr4IMKHSNWCiUQ6AEIRzAE#v=onepage&q=immunotherapy%20visualization&f=false

[5]https://physics.cancer.gov/research/2012/june/po_news_f.aspx

[6]http://cancerres.aacrjournals.org/content/62/20/5785.full

[7]http://www.math.pku.edu.cn/teachers/yaoy/math112230/Lecture20_YuanXR_visualization.pdf

[8]https://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf

[9]http://ieg.ifs.tuwien.ac.at/~aigner/teaching/ws04/infovis_ue/aufgabe3/references/tominski03axes.pdf

[10]http://www.informatik.uni-rostock.de/~hs162/pdf/infovis05

[11]http://people.cs.vt.edu/~north/infoviz/mgv.pdf

[12]http://www.math.pku.edu.cn/teachers/yaoy/math112230/Lecture20_YuanXR_visualization.pdf

[13]http://www.math.le.ac.uk/people/ag153/homepage/LStuart.pdf

[14]http://www.cc.gatech.edu/~stasko/papers/avi14-touch.pdf

[15]http://ovii.oerc.ox.ac.uk/sp-pcp/assets/data/thesis.pdf