Clustering Using D3

Tejas Karangale
Master of Science in Information Management
University of Washington
Seattle, United States
tejk@uw.edu

Scarlett Dias
Master of Science in Information Management
University of Washington
Seattle, United States
sdias@uw.edu

Vaisakhi Mishra
Master of Science in Information Management
University of Washington
Seattle, United States
vm11@uw.edu

***Abstract*— K-means Clustering is one of the simplest unsupervised algorithm used to cluster unlabeled data. It is usually one of the first clustering algorithms taught to a student wishing to learn machine learning algorithms. The purpose of creating this interactive tutorial was to teach the function of k-means algorithm in a more visual and intuitive way, rather than the traditional textual way. Our Martini glass structured tutorial first gives the reader, a brief step by step and easy to understand tutorial and then allows the user to experiment with the parameters used in k-means clustering algorithm.**

Keywords—k-means, clustering, algorithm, interactive, tutorial, D3, visualization

I. INTRODUCTION

Suppose you own a rental company and have thousands of customers. You want to improve your sales and create a business strategy, but every customer is different and it isn't feasible for you to create a different strategy for every customer. Here is when clustering comes into picture. Customers can be clustered in different groups, for example: 10 groups and you can easily create 10 different business strategies for every customer cluster.

Clustering is an unsupervised machine learning algorithm which divides the data points into logical groups such that data points in the same group are similar to each other than the points in other groups. K-means is a versatile algorithm and can be used for any type of grouping. We plot the mean distance to the centroid as a function of K and the "elbow point," where the rate of decrease sharply shifts is used to determine the value of K. [1] This algorithm is really useful to get a logical structure from our random data.

II. RELATED WORK

We looked at various tutorials available that teach the algorithm online. The current resources are in the form of websites or info visuals or videos. We found that most of them come under two distinct categories.

A. *Static Descriptions*

Some tutorials look at explaining k-means clustering using a mixture of text and static visuals. The website tutorial on "A Tutorial on Clustering Algorithms" seeks to explain the algorithm with lots of text and formulae on the page with minimal visuals. The website "Big Data Made Simple" on the other hand does try to make it easier with static visuals.

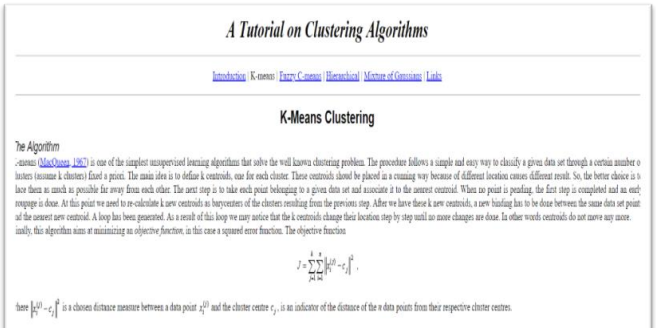

Fig. 1. Tutorial on Clustering Algorithms – Text Based [2]

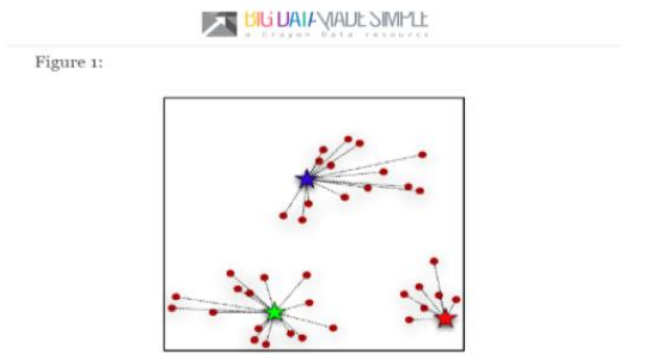

Fig. 2. Big Data Made Simple – Static visualization [3]

B. *Interactive Tools*

There are a few Java or Flash based tools available that allow the user to play with data and interact with them. These tools however do not give the user much information in the way of a step by step explanation. This can be daunting to a person who does not know the basics of the algorithm.

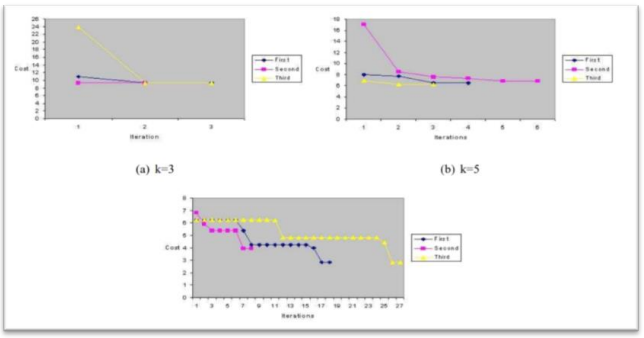

Fig. 3. Interactive Tool to Understand K-Means [4]

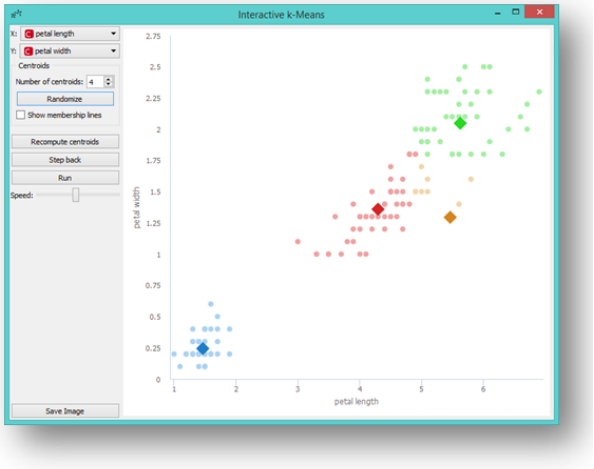

Fig. 4. Education Widget to learn K-Means [5]

III. METHODS

A. *Martini Glass Structure*

The Martini Glass Structure is a common way of storytelling. It consists of three parts. The starting point (the base of the glass) gives the reader an overview of the narrative. The Author-driven stage usually narrows in and focuses on the core explanation of the story. The User-driven stage is usually interactive and allows the user to explore, turning the author narrative into a user driven one.

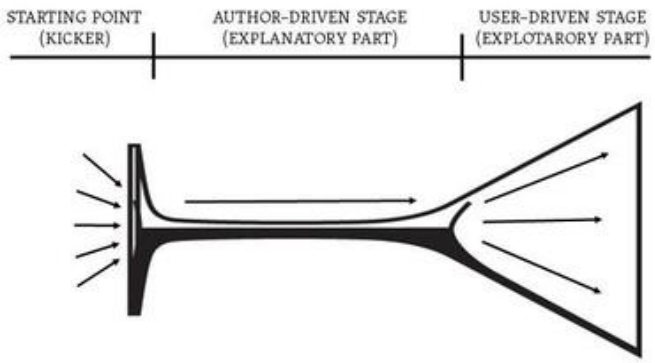

Fig. 5. Martini-Glass Exploratory Sturcture

Since we wanted to implement an exploratory style in our tutorial, we decided to create a web-page that followed this style and structure. We started with a basic overview of what K-means clustering is and when it is used. We then proceeded to create a step-by-step animated tutorial that allowed the user to understand the algorithm. The user could scroll forwards or backwards to repeat the animate at any step and understand it better. After the explanation of the algorithm, we tied the end results with one of the scenarios where the algorithm could be used.

We finally created an interactive tool that the users could play around with and look at how the clusters change depending on the number of data points and clusters specified. Using the buttons provided, the user could have control over what they wanted to focus on.

B. *The Explanatory Stage*

We started off the tutorial by plotting randomly generated 25 data points in our vector space. These data points were generated in the range of the svg's height and width and were displayed on the page as a scatter plot.

Fig. 6. Scatter Plot of Random Data Points

On scroll down, we implemented the next step. The second step of our tutorial consists of randomly placing three centroids (k = 3) in the vector space. We have set the shape of the centroid as a diamond (square rotated by 45°) and colored the three centroids differently so that they're distinguishable from the data points in the first step.

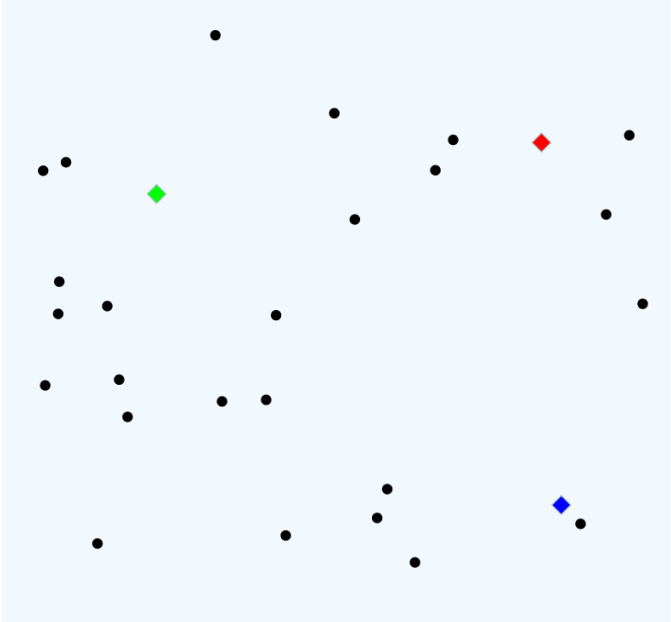

Fig. 7. Plot the Centroids

After plotting the centroids in the previous step, we then calculate the Eucliden distance of all the data points from all three centroids and assign data points to their respective clusters by considering the lowest Euclidean distance from the centroid to the data point. We show the assignment by coloring the data points to the same color as its centroid point. This forms the initial clusters in the data set. Along with that, in the same section we explain that the clusters could be made by using two types of distances – Euclidean and Manhatan and discuss the scenarios when one could be used.

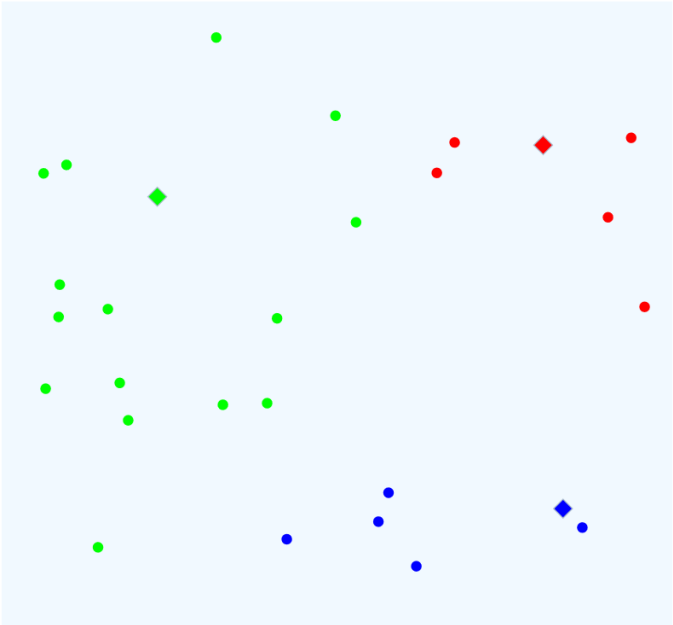

Fig. 8. Cluster Assignment

In the next step, we show the Euclidean distance from the data point to its cluster centroid using a hover tool tip. We draw lines from the centroids to the cluster points and use length encoding for the Euclidean Distance. However, the tool tip that shows the Euclidean distance is triggered on hover over the line and well as the point with which the line is associated. During this step, we also explain the statistical formula of Euclidean distance.

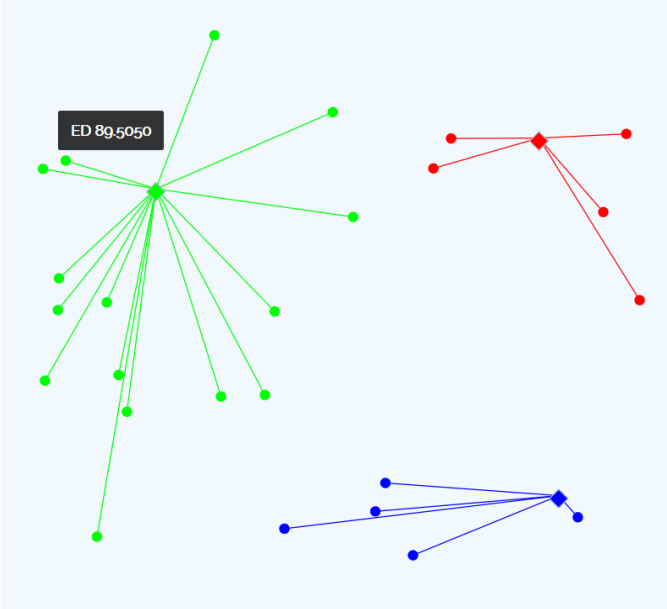

Fig. 9. Euclidean Distance

The next step is an animated graph of every subsequent iteration of the previous steps. In the animation loop, we first update each cluster centroid to the mean of all the points currently present in the cluster and show its transition on the graph. Meanwhile we also indicate that centroids have changed, using text. Next we re-assign the clustr points by using Euclidean distance and show the change of the points by first changing the colour of the points, then changing their respective lines. During this step, we indicate the total number of points that changed their clusters in that iteration. This loop comtinues till the local optimum of k-means clutering aka the condition when no points change their clusters is reached.

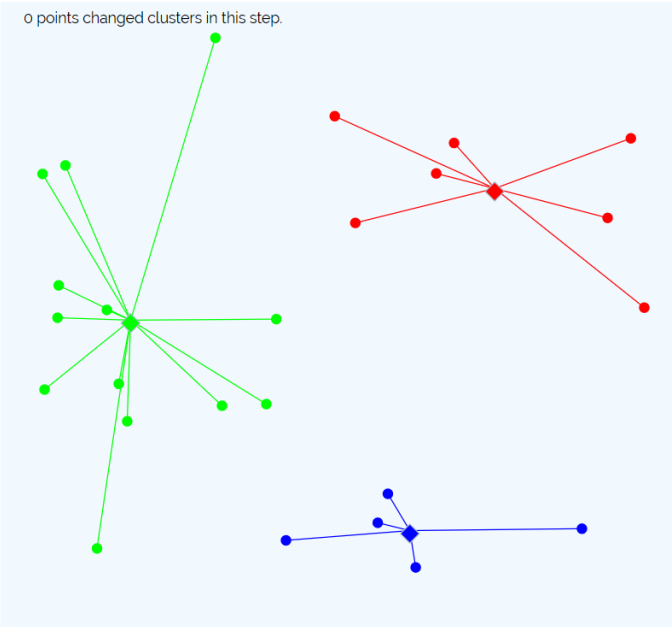

Fig. 10. Animated Clustering till Local Optimum reached

C. *The Exploratory Stage*

After the end of the explaination of the algorithm, we let the users try out the k-means clustering algoritn with an interactive graph. Here, were have 30 random points (N) initally plotted in a scatterplot with 3 default clusters (K) at start, but give users the liberty of entering upto 1500 nodes for clustering and increase or decrease the number of clusters in the range of 1-15.

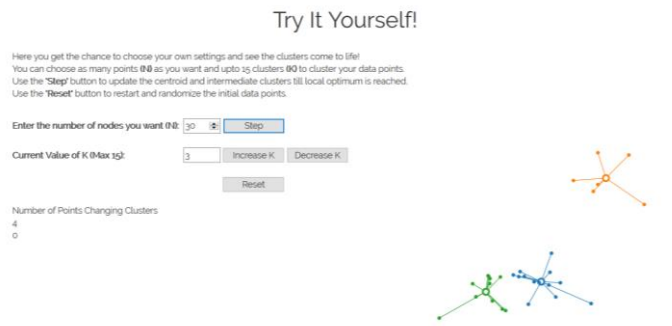

Fig. 11. Interactive Tool

We use similar visual encoding in both the tutorial and the interactive part, to keep the algorithm semantics consistent. Also, in this interactive section, we let the user step through the graph for the iterations by using a step button, all the while showing the number of points changing the clusters in that particular step. User can do this till the local optimum is reached and then chose to either change the number of nodes or clusters or reset the whole graph to default by using the reset button. In the interactive graph, we chose to not show the Eulidean distance on hover as the lines and points would get really close, and might overlap when N is increased.

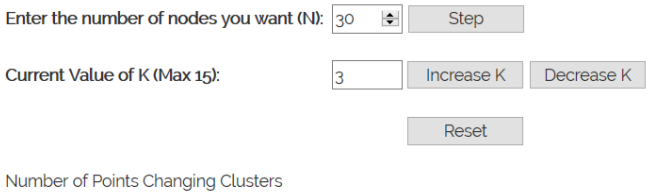

Fig. 12. Interactive buttons for the user to test

IV. Results

The tutorial implemented in D3.js allows us to have a lot of control over the visual elements. We can add custom data points according to our requirement. The web page starts with a simplified introduction of what the k-means algorithm is. The scroller visualization then helps the user understand k-means clustering step by step. The main advantage of using a scroller visualization is that, the user can concentrate on each step of the algorithm one step at a time. The smooth scrolling allows easy transition and the user can understand the process flow of the algorithm. The user is informed about the drawback the algorithm has and the optimum step at which the clustering should stop.

The interactive visualization at the end, gives user a chance to play around with random number of points of his choice and also the number of clusters he wants to create. Instructions on how to use the simple interaction are provided. The user can easily view each step of each cluster formation by clicking the corresponding buttons.

V. Discussion

Machine learning is a booming field currently and to educate people with basic machine learning algorithms in a nontraditional approach was our goal. K-means clustering algorithm is an unsupervised machine learning algorithm which in not memory heavy and easy for the computer to run. It's cost in terms of memory usage is very less so we finalized this algorithm for our interactive tutorial.

After analyzing existing tutorials on k-means clustering, we observed that these tutorials were very static, they didn't have adequate interactions for a user to experiment or the tutorials just did not have adequate information about the process. One of the most important sub-problems we faced was whether to just create the interactive visualization with a text explanation at the start or to create a narrative visualization and then allowing the user the interact. After brainstorming, we decided to go with the Martini Glass Structure as we wanted to include the author-driven explanatory part so that the user is crystal clear about the concepts used in k-means clustering algorithm.

VI. Future Work

Our next plan is to teach the user using a sample dataset. The user will be given a choice of three datasets and he can see the working of k-means clustering algorithm on the dataset itself – e.g. We will have a sample dataset of songs which has genre as a feature. The user can then see how the songs are clustered according to the genres and the number of clusters (k) specified.

We also plan to include a short quiz at the end of the explanatory part where we can ask a few conceptual questions to the user. We also plan to add an option to toggle between Euclidean and Manhattan distance. We will show a side by side comparison of how the clustering patterns change on the usage of two different types of distance metrics.

VII. Acknowledgement

The tutorial scroller was implemented by using graph-scroller.js. The implementation of K-means algorithm was inspired by Tech-Ni Blog's implementation and Joe Beuckman's implementation of K-Means clustering.

VIII. References

[1] datascience.com, "Introduction to K-means Clustering," 06 12 2016. [Online]. Available: https://www.datascience.com/blog/k-means-clustering.

[2] "A Tutorial on Clustering Algorithms," [Online]. Available: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html.

[3] Jeevan, Manu;, "Possibly the simplest way to explain K-Means algorithm," [Online]. Available: http://bigdata-madesimple.com/possibly-the-simplest-way-to-explain-k-means-algorithm/.

[4] P. Goffin, "Interactive Tutorial for the Understanding of the K-Means Algorithm," Doctoral dissertation, University of Leeds, School of Computer Studies, 2008.

[5] "Interactive k-means," [Online]. Available: https://orange3-educational.readthedocs.io/en/latest/widgets/kmeans.html.