ISSN: 0975-766X
CODEN: IJPTFI
*Available Online through* *Research Article*
**www.ijptonline.com**

# DATA VISUALIZATIONTECHNIQUE TO VISUALIZE CRIME DATA USING CHERNOFF FACES

**Ankita Goyal , Priyanshi Gupta, Arushi Gupta, Jabanjalin Hilda J, Gladys Gnana Kiruba B**[1,2,3,4,5]
[1,2,3,4,5]School of Computer Science and Engineering, VIT University, Vellore, TamilNadu.
*Email: jabanjalin.hilda@vit.ac.in*

*Received on: 15-02-2017* *Accepted on: 27-03-2017*

**Abstract:**

**Background and Objective:** In the present scenario when there is so much of data generated from every field, handling and extracting the correct information from this data is a cumbersome task. The main task while extracting information is to understand the dataset given to you and differentiate the various values. This paper will use a crime dataset of 50 states of United States of America to implement chernoff faces which is used to represent data in form of various types of faces. **Methods:** This dataset is analyzed using R tool inbuilt functions and in order to visualize the dataset properly the parameters of the built in functions were changed accordingly. The same dataset was visualized using K-means algorithm in Rapid Miner. **Results:** Using this technique 6 cluster are formed and plotted in a line graph.It has been observed that to understand data in more depth and detail, visualization using chernoff faces is a better option than clustering using K-means. With the help of this pictorial representation the user can notice small details about data which is not usually observable when statistical methods are used.

**Keywords:** Chernoff faces, Clustering, K-means algorithm, Data visualization, Crime

## 1. Introduction

One of the main drawbacks in information exploratory is the growing quantity of information that has to be understood and then worked upon. The actions in various fields including commercial, engineering and administration, produce a huge quantity of data. This information is accumulated in very large databases because it is crucial and also mining this information will produce results which can be productive in various fields. Mining therefore required correct and normalized information, and thus, is anirk some job. Finding useful data in a actual great repository with loads of information substances stays an exhaustive and slow task, even with the most advanced data analysis systems. The

process cannot be completely automated and at some point human creative intelligence has to be involved as they are the one who will be designing algorithms according to the requirements of the dataset and thus it cannot be completely replaced by computers. Humans will therefore remain to be a crucial part of search and analyzes of the data. While managing vast datasets, in any case, people should be adequately upheld by the computer. One often received method for supporting the human in investigating expansive measure of information is to envision the data. Visualization of data backs to the time even when automated techniques by computers were not used to create visualization of the datasets which have distinctive many dimensional semantics. Since PCs are utilized to make perceptions, numerous remarkable visualization procedures have been created and the current ones have been unpremeditated to be utilized for substantial information sets and make the displays significantly more intelligent. Various methods have been developed to visualize data which has several attributes. There are various statistical techniques like:

- Tree Maps
- Scatter Plot
- Bar Graphs
- Scatter Plot
- HeatMaps
- Network
- Histogram
- Gantt Chart

Though these techniques are easy to implement and visualize but the main drawback of these statistical methods is that they are not the proper way to visualize and understand things in detail. Users can not perform proper evaluation and thus cannot jump to correct conclusions regarding their dataset. A more detailed and in depth method of visualizing data is to use related symbols and metaphors to represent general information. As an example to represent stock data an image metaphor of "money tree" has been used[1]. In the paper we have shown a more detailed method of visualizing dataset consisting of large numbers of attributes. It is a graphical method where each attribute of our dataset is mapped to a feature of face like width of smile, height of face, width of ear etc. This technique is given the name of chernoff faces. The main reason behind using this method over all the implementation tool is because faces enlighten emotional

responses from the user. Since every individual does a good study of their own faces it is easy for them to notice tiny differences between each and every face and thus can give a proper interpretation of their problems. The paper is organized as follows. The implementation of chernoff faces has been explained in the beginning, it describes the tool we have used and the dataset we worked on. This is followed by introduction of other visualization techniques done by various other authors. Implementation of the same dataset is done by rapid miner tool by making use of K-means algorithm which gives scatter plot as result. In the end comparison has been made between the two techniques implemented.

## 2. Related Work

This is one of the major approaches that can be used to represent a dataset. The approach has been used by **Lindberg** is as follows: He accumulated the facial and prominent features of the face like nose, hair, mustache into a photograph. Each feature corresponds to a numerical value which helps in distinguishing different data in a dataset.[2]

Another method has been given by **Anderson** which is developed by using "glyphs" .Glyphs are cluster of visual objects, these objects have rays of different lengths and directions which extend away fromthe boundary. The length of every ray represents the value of a variable.[3]

**Pickett and White** used triangles to represent 4 variable by measuring length of three sides and orientation. These measures help to portray the given dataset. Another technique used to visualize data of various attribute is called *pixel-oriented technique*. The idea driving pixel-oriented procedures is to coordinate every information esteem to a pixel and delineate the information values having a place with one variable in isolated windows. Fundamentally this procedure utilizes just a single pixel for each data esteem, therefore the space and memory consumption is less and thus we are able to visualize a large dataset. Since every information esteem is spoken to by one pixel, the fundamental issue is the way the pixels are to be masterminded on the screen. This pixel-oriented methods utilize diverse type of provisions for various objectives. In the event that a user needs to represent a colossal dataset, the user can utilize a perception procedure which is not reliant on any user particular question and it can mastermind the information as indicated by some factor and uses a screen-filling pattern to organize the information values on the show. On the contrary, if there is no ordering of the data and we have to visualize an interactive database display, we can get it by executing a query. Instead of coordinating an information esteem straightforwardly with a shading, the question subordinate visualization

strategies compute the separations amongst information and query values, consolidate the separations for every information into a overall separation, visualize the separations for the factors and the general separation is sorted as for the overall separation. In icon based visualization, we have analyzed how we can use metaphor-based icons instead of abstract icons in icon-based visualization of a multivariate data to create mosaic representing that data.[3] The development of an Image Mosaic comprises of the accompanying steps: First we chose images which we can use as mosaic tiles, then we pick a grid, so that we can arrange mosaic tiles in the grid and execute shading correction on the tiles to coordinate the target picture. The information set that we have taken for instance that portrays an extreme drought that happened in the semi-arid-dry upper east of Brazil. The information set comprises of 6 parameters and a cluster ID at each 2D position making it a sum of 7 parameters. Sound maize collect are delineated by a thick, yellow cob, while ailing maize are spoken to by a thin, cocoa cob. Since we have to represent data values for six parameters, we subdivide the maize cob picture into six locales, and assemble the subsequent picture for every one of these parameters and the combination depends on six picture parts. The cluster IDs are depicted by the background colors[4]. After the mosaic tiles have been outlined, we find a format for the tiles and place them as indicated by that. 3 different types of grid layout are: scattered formats, regular formats and multi-resolution formats. For Image Mosaics we adopt scattered layout. One of the drawback with this approach is covering of icons in districts which have high station thickness. To beat this disadvantage, we can go for consistent and multi-resolution designs. Regular designs take the average of values of all measurement stations and display it in each grid cell. In a multi-resolution design, symbols are put in a locale, if we find only one cluster in that locale. If not, the area is again subdivided[5]. At that point, parameter estimations of comparative stations from a similar group are amassed and in this way we can sidestep merging of locale which has distinctive properties. After the design step has been played out the following undertaking is to choose which symbols must be set on the guide. This choice in light of design, particularly attributable to various area merging procedures[6]. The last stride is to, parameterize the figurative icons, by applying a color correction. Color choice for the parameter classifications for each of the six symbol parts, while planning the symbols, can be viewed as a sort of color correction with just three hues.

3. **Methodology:** The main intention of Chernoff faces is to represent various variables of dataset provided using facial expressions, every facial expression will be mapped to a column or attribute. To implement this we have taken a dataset

of crime rate in US states. The dataset consists of 50 states and 8 columns which represent eight different types of crime happening in US states[7]. The following figure 1 displays the dataset which will be implemented: To implement Chernoff faces we have made the use of **R** tool.

| state | murder | forcible_rape | robbery | aggravated_assault | burglary | larceny_theft | motor_vehicle_theft |
|---|---|---|---|---|---|---|---|
| United States | 5.6 | 31.7 | 140.7 | 291.1 | 726.7 | 2286.3 | 416.7 |
| Alabama | 8.2 | 34.3 | 141.4 | 247.8 | 953.8 | 2650.0 | 288.3 |
| Alaska | 4.8 | 81.1 | 80.9 | 465.1 | 622.5 | 2599.1 | 391.0 |
| Arizona | 7.5 | 33.8 | 144.4 | 327.4 | 948.4 | 2965.2 | 924.4 |
| Arkansas | 6.7 | 42.9 | 91.1 | 386.8 | 1084.6 | 2711.2 | 262.1 |
| California | 6.9 | 26.0 | 176.1 | 317.3 | 693.3 | 1916.5 | 712.8 |
| Colorado | 3.7 | 43.4 | 84.6 | 264.7 | 744.8 | 2735.2 | 559.5 |
| Connecticut | 2.9 | 20.0 | 113.0 | 138.6 | 437.1 | 1824.1 | 296.8 |
| Delaware | 4.4 | 44.7 | 154.8 | 428.2 | 688.9 | 2144.0 | 278.5 |
| District of Columbia | 35.4 | 30.2 | 672.1 | 721.3 | 649.7 | 2694.9 | 1402.3 |
| Florida | 5.0 | 37.1 | 169.4 | 496.6 | 926.3 | 2658.3 | 423.3 |
| Georgia | 6.2 | 23.6 | 154.8 | 264.3 | 931.0 | 2751.1 | 490.2 |
| Hawaii | 1.9 | 26.9 | 78.5 | 147.8 | 767.9 | 3308.4 | 716.4 |
| Idaho | 2.4 | 40.4 | 18.6 | 195.4 | 564.4 | 1931.7 | 201.8 |
| Illinois | 6.0 | 33.7 | 181.7 | 330.2 | 606.9 | 2164.8 | 308.6 |
| Indiana | 5.7 | 29.6 | 108.6 | 179.9 | 697.6 | 2412.0 | 346.7 |
| Iowa | 1.3 | 27.9 | 38.9 | 223.3 | 606.4 | 2042.7 | 184.6 |
| Kansas | 3.7 | 38.4 | 65.3 | 280.0 | 689.2 | 2758.1 | 339.6 |
| Kentucky | 4.6 | 34.0 | 88.4 | 139.8 | 634.0 | 1685.8 | 210.8 |
| Louisiana | 9.9 | 31.4 | 118.0 | 435.1 | 870.6 | 2494.5 | 318.1 |
| Maine | 1.4 | 24.7 | 24.4 | 61.7 | 478.5 | 1832.6 | 102.0 |
| Maryland | 9.9 | 22.6 | 256.7 | 413.8 | 641.4 | 2294.3 | 608.4 |
| Massachusetts | 2.7 | 27.1 | 119.0 | 308.1 | 541.1 | 1527.4 | 295.1 |
| Michigan | 6.1 | 51.3 | 131.8 | 362.9 | 696.8 | 1917.8 | 476.5 |
| Minnesota | 2.2 | 44.0 | 92.0 | 158.7 | 578.9 | 2226.9 | 278.2 |
| Mississippi | 7.3 | 39.3 | 82.3 | 149.4 | 919.7 | 2083.9 | 256.5 |
| Missouri | 6.9 | 28.0 | 124.1 | 366.4 | 738.3 | 2746.2 | 443.1 |
| Montana | 1.9 | 32.2 | 18.9 | 228.5 | 389.2 | 2543.0 | 210.7 |
| Nebraska | 2.5 | 32.9 | 59.1 | 192.5 | 532.4 | 2574.3 | 316.5 |
| Nevada | 8.5 | 42.1 | 194.7 | 361.5 | 972.4 | 2153.9 | 1115.2 |
| New Hampshire | 1.4 | 30.9 | 27.4 | 72.3 | 317.0 | 1377.3 | 102.1 |
| New Jersey | 4.8 | 13.9 | 151.6 | 184.4 | 447.1 | 1568.4 | 317.5 |
| New Mexico | 7.4 | 54.1 | 98.7 | 541.9 | 1093.9 | 2639.9 | 414.5 |

**Figure 1: Crime Data set of United States of America.**

**Using R**

R is an open source programming tool which supports big data analytics and graphics. In the following implementation we have used R to combine a large data set and visualize the whole data set using graphical functions provided in R. After the visualization of the whole data set it will be easy to provide results regarding the dataset.R has many inbuilt packages and libraries with proper documentation. We can study the documentation of the function we want to make changes to and change our code accordingly by making changes to the parameters of that particular function. The documentation can be studied by adding '?' symbol in front of the built-in function we want to use for implementation of the given dataset.

**4. Implementation**

The first step is to import the dataset in the csv(comma separated values) format. After loading the library we can start making faces by using the faces() function. The first important parameter of this function will be the name of the dataset. By default colored faces will be displayed, a screenshot of which is displayed below:

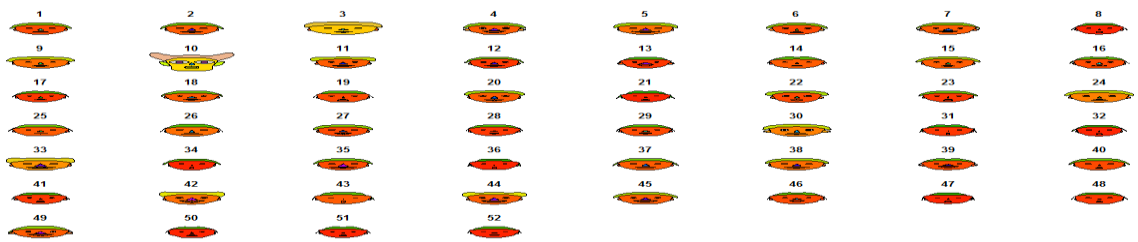


**Figure 2: Dataset being converted to chernoff faces.**

We can see that all the faces have smiles which are not a good feature for our dataset as it is a crime dataset, to remove the smile feature we can make changes to columns of this matrix and store the values in new matrix. After removing the smiles and converting the numbers to state names our dataset will look like:


**Figure3: Removing smiles from the crime data set and giving each face state name.**

To change these by default faces we can change these shapes to santa clause faces by using the function faces(crime[,2:8], face. type=2),now these faces will look like:

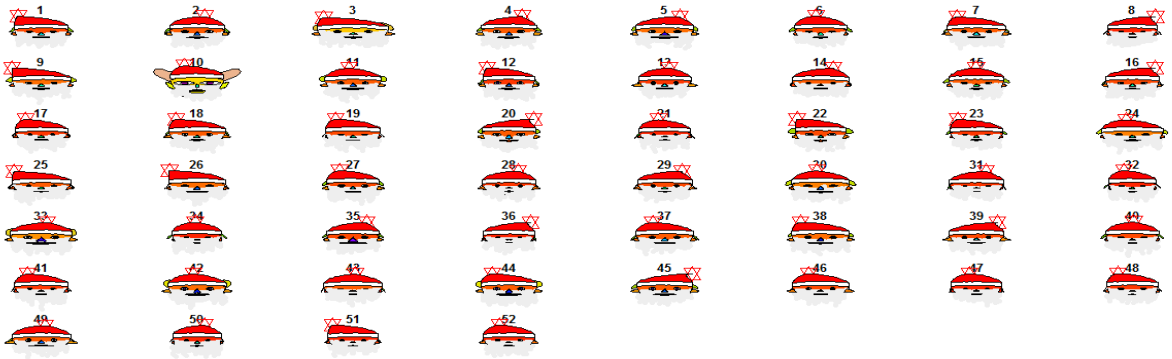

**Figure 4: Different visualization of the same dataset**

In the dataset every crime is associated with a facial feature.

## 5. Comparison

For comparison we have used another technique , K-means algorithm for implementation and visualization of the same data set.

**K-means algorithm**.

K-means clustering is one of the strategy for group investigation which plans to partition n perceptions into k bunches in which every perception has a place with the group with the closest mean.

**Procedure**

1. At first, the number of groups must be known given it a chance to be k

2. Choose an arrangement of K examples as centers of the bunches.

3. Next, the calculation considers every example and relegates it to the bunch which is nearest.

4. The bunch centroids are recalculated either after entire cycle of re-task or every occurrence task.

5. This procedure is iterated.

K means calculation complexity is O(tkn), where n is instances, c is bunches, and t is iterations and moderately effective. It frequently ends at a local optimum. Its weakness is pertinent just when mean is characterized and need to indicate c, the number of bunches, ahead of time. It is not able to handle noisy information and exceptions and not appropriate to find groups with non-convex shapes [8].

We have used the Rapid miner tool for the following implementation. Rapid Miner is an integrated environment studio using which we can easily implement data analytics techniques such as machine learning, fuzzy logic, artificial intelligence, neural networks, supervised and non-supervised learning[9].

To analyze our dataset we have applied K-means algorithm to form clusters of the states. Each state is given a number and according fall in a particular cluster. The result of five clusters, their visualization result is as shown below. The x axis represents numbers and the y axis corresponds to the crime.

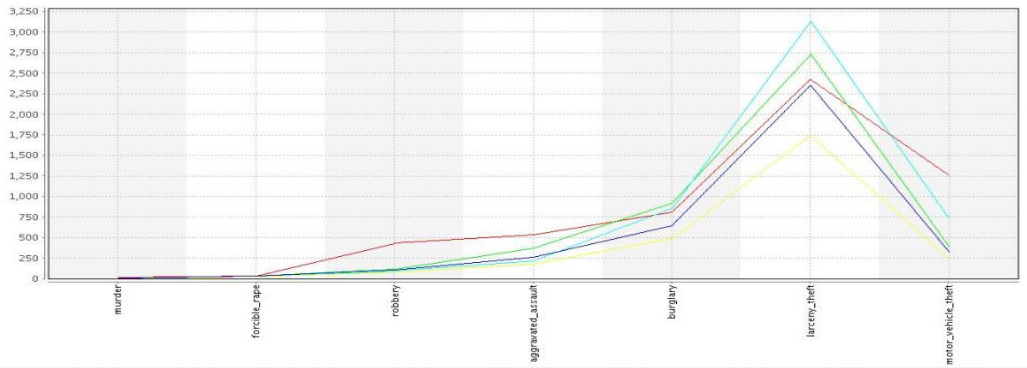


**Figure 5: Results obtained after visualization in RapidMiner**

The x-axis represents the crimes and the y-axis represents centroid of each attribute. The lines are plotted in different colors where each line corresponds to each cluster. To measure the intensity of a crime we can read the corresponding x and y axis. If we take the cluster_3 from the result obtained from rapid miner and compare them with the chernoff faces we can observe that all the numbers in cluster_3 also have similar chernoff faces.

## 6. Conclusion

Throughout their lives human beings have grown up reacting to facial expressions and emotions and thus visualization using chernoff faces will produce a deeper understanding of multivariate data. Significant and different features of faces are quickly observed and also easy to memorize. K-means calculation does not function admirably with clusters of various size and diverse thickness. On the contrary, with the help of Chernoff faces we can easily distinguish each state and assign it to a particular cluster without the involvement of any kind of numerical data.

## 7. References

1. Eades P, Shen X. MoneyTree: ambient information visualization of financial data. InProceedings of the Pan-Sydney area workshop on Visual information processing 2004 Jun 1 (pp. 15-18). Australian Computer Society, Inc..
2. Lindberg, D.A.B., Unpublished communication.
3. Nocke T, Schlechtweg S, Schumann H. Icon-based visualization using mosaic metaphors. InNinth International Conference on Information Visualisation (IV'05) 2005 Jul 6 (pp. 103-109). IEEE
4. Peng W, Ward MO, Rundensteiner EA. Clutter reduction in multi-dimensional data visualization using dimension reordering. InInformation Visualization, 2004. INFOVIS 2004. IEEE Symposium on 2004 Oct 10 (pp. 89-96). IEEE.
5. McKenna T, Arce GR. New image mosaic structures. Technical report, Department of Electrical and Computer Engineering, University of Delaware; 2000.
6. Nocke T, Schumann H, Böhm U. Methods for the visualization of clustered climate data. Computational Statistics. 2004 Feb 1;19(1):75-94.
7. Adam Finkelstein and Marisa Range. Image mosaics. Technical Report TR-574-98, Princeton University, Computer Science Department, March 1998.
8. Andrews DF. Plots of high-dimensional data. Biometrics. 1972 Mar 1:125-36.
9. McCallum A, Nigam K, Ungar LH. Efficient clustering of high-dimensional data sets with application to reference matching. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining 2000 Aug 1 (pp. 169-178). ACM.