CLUSTERING TECHNIQUES TO ANALYZE SKIN CANCER

Dr.S. Gomathi[1]
Assistant Professor
Department of Computer Science (PG)
PSGR Krishnammal College for Women
Mail Id:mailtogomathisrinivasan@gmail.com

Ms. J. Angel Monica[2]
M.Sc., Data Analytics
Department of Computer Science (PG)
PSGR Krishnammal College for Women
Mail Id:jangelmonica18@gmail.com

Ms. R. Siva Priya[3]
M.Sc., Data Analytics
Department of Computer Science (PG)
PSGR Krishnammal College for Women
Mail Id:siva3priya1999@gmail.com

Abstract:

This paper presents a cluster analysis of skin cancer dataset using orange data mining tool, open source data visualization software. Using clustering techniques in this software we analyze and extract the information from data. Clustering is one of the analytical methods which include the distribution of data into groups of identical objects. Clustering algorithms namely Hierarchical, K-Means and silhouette have been used tocluster the data based on the information of the gender (sex). In this paper using various widgets in orange tool we implemented the dataset about the people who are affected by the common pigmented skin lesions.The analyzed data involve lesion id, image id, diagnosis (dx) and how the diagnosis was made (dx type) of men and women.

Index Terms: Hierarchical, K-Means, Silhouette Clustering, skin cancer, analysis, orange tool.

Introduction:

In this paper we used orange data mining for clustering. Orange tool is an analyzing tool which is an open-source software package.It is a part of machine learning. Clustering is also known as cluster analysis and that is an important subject in data mining. Data mining used to extract the useful information from the large amounts of data. It helps to make better decisions with the fastest technique and that is used to analyze the huge amount of data in less time effortlessly.This tool is used when we have large amount of data in sectors (For e.g.: Business, Healthcare, marketing organizations, etc..,)The most important Data mining techniques are classification, clustering, Regression, association rules, prediction, sequential patterns and outer detection.

We used clustering for skin cancer dataset. A cluster is composed of a number of related objects collected and grouped together. Clustering is used in many fields like machine learning, graphics, etc.., Clustering is used to solve the business problems. It contains all objects in the data set. We used Centroid-based clustering, which is represented by a central vector. [1]

Dataset: Skin cancer data has been collected in MS-Excel.

Types of clustering methods are given below:

A) Hierarchical Clustering
B) K means
C) Silhouette Clustering

Upload file by the following steps:

1. Click file→Choose a data→Click Reload.
2. The data will be uploaded in a file widget.

HIERARCHICAL CLUSTERING OR HIERARCHICAL CLUSTER ANALYSIS:

Hierarchical clustering is also known as Hierarchical cluster analysis or HCA. HCA is a process of cluster analysis which builds a hierarchy of clusters. Hierarchical clustering is performed with either raw data or distance matrix. Here we used a raw dataset. If the raw data is provided, then the software can able to compute a distance in a background. Hierarchical clustering works by treating each data as a separate cluster.The distance between two clusters has been computed based on gender.

This analysis is used to find out the distance between rows. [2]

1. Upload data in a file widget.
2. Choose a distances widget to compute distance between rows.
3. Click on hierarchical clustering widget and select the particular cluster in a widget. (fig.1)

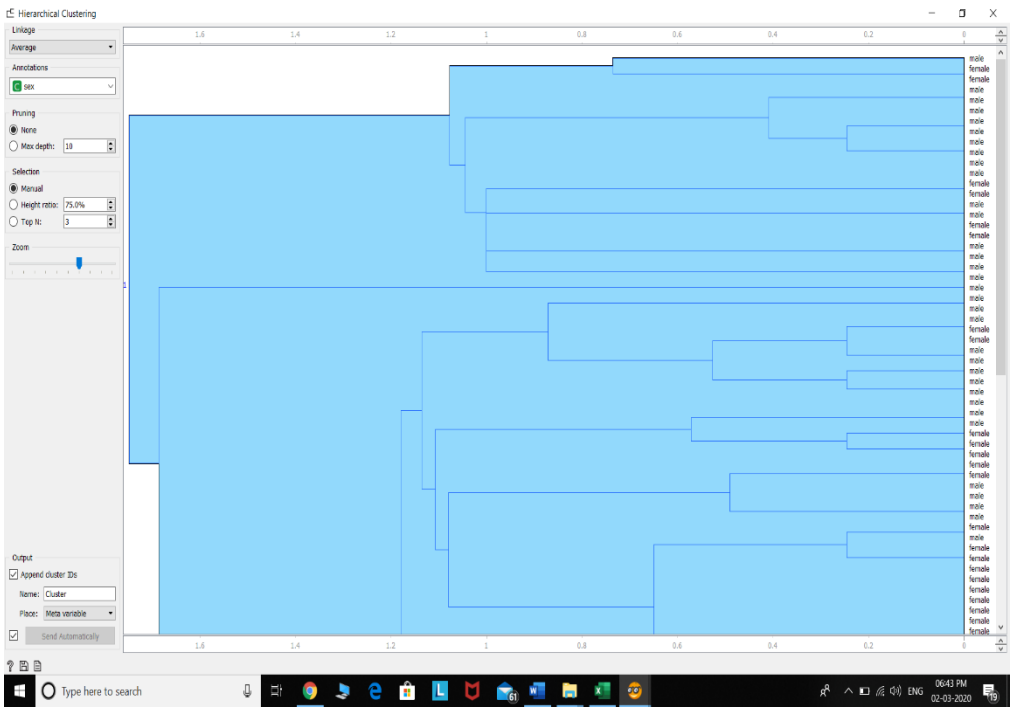


Fig.1 Particular cluster has been selected

4. The outcome of the selected cluster is represented in a boxplot[fig.2]

Fig.2 Workflow of box plot

5. From this box plot (Fig.2), we could predict that how many male and female affected by skin cancer in different localization.

K MEANS CLUSTERING:

This k means method is used to specify the number of clusters. K mean is a partition method. [3]K-means is a distance-based algorithm or a centroid-based algorithm where we calculate the distances to allocate a point to a cluster.K-Means clustering is used in a multiplicity of examples in real life like Education, Diagnostic systems, Wireless sensor networks, etc.,

Here we are given a data set of skin cancer, with certain attributes and values for those attributes. The target is to categorize those items into groups. To attain this we used the k-Means algorithm, an unsupervised learning algorithm.The algorithm will categorize the data into k groups of similarity.

We can find out the number of clusters by the following steps:

1. Choose a Paint data widget
2. Paint some data in a widget
3. Click K mean widget and decide how many clusters do we need and fix values in fixed radio button.
4. Here we choose k values as 3.
5. Choose Scatter plot widget.
6. The given number of clusters will be displayed in a scatter plot.

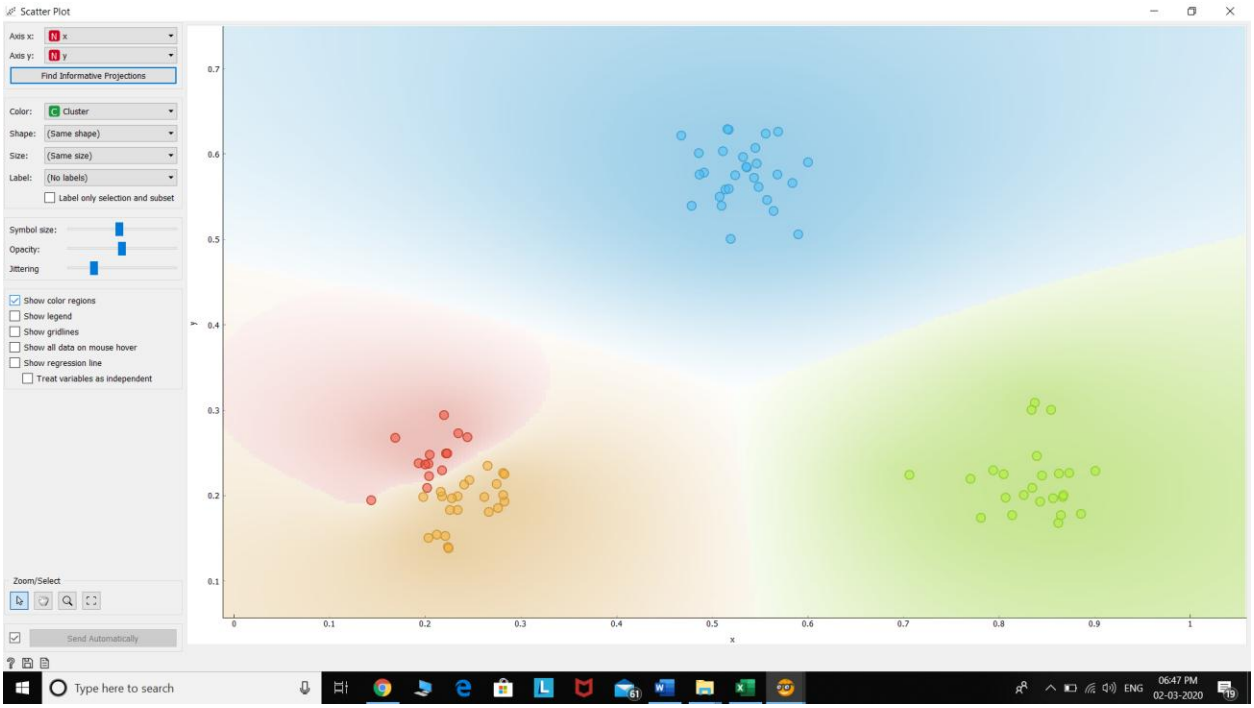


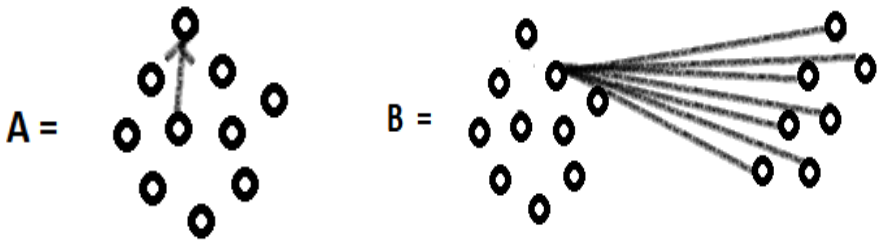
Fig.3 Workflow of scatter plot

SILHOUETTE CLUSTERING:

This silhouette clustering is used to find out the distance of the each point in a cluster or to find out the distance between each cluster.[2]

To find out the distance between the clusters:

=B-A / MAX(A,B)



1. Choose paint data widget.
2. Paint some data in a given dialog box.
3. Choose number of clusters in k means widget.
4. Here we choose the k value as 3.
5. Click on silhouette plot widget.
6. This widget shows the data in a graphical representation within the clusters using Euclidean distance**[fig.4]**
7. Select the maximum test score in the silhouette plot.

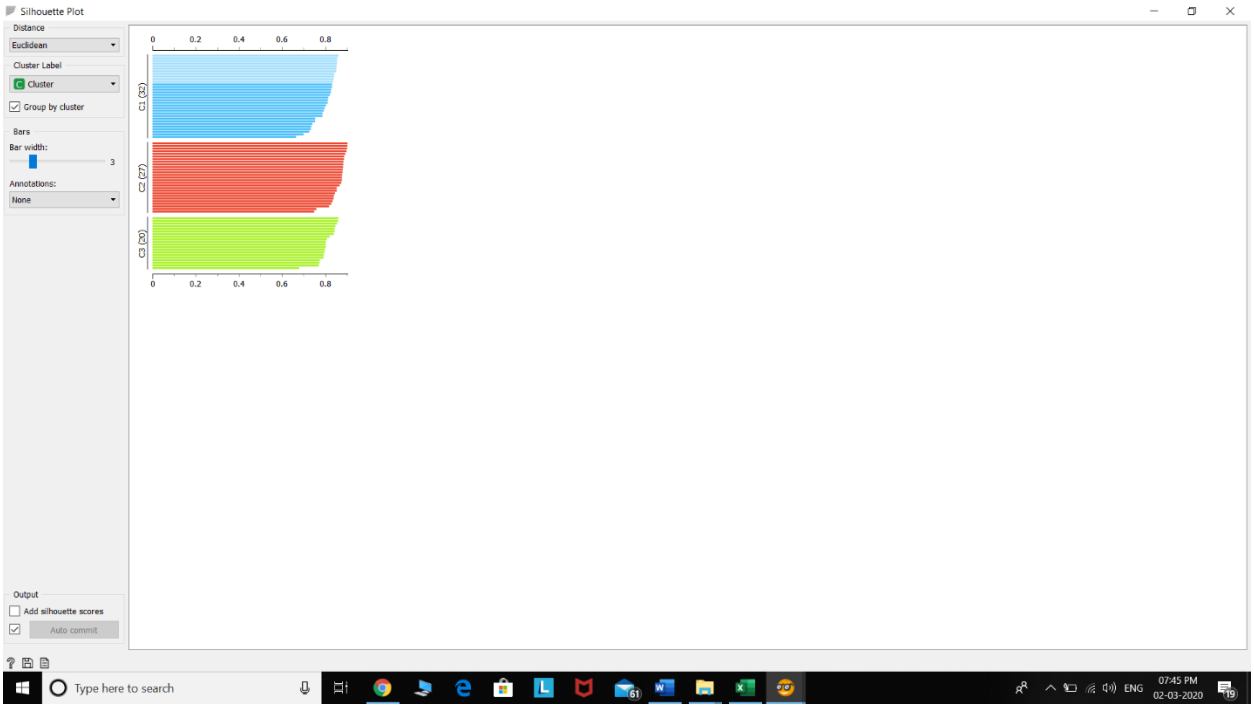


Fig.4 Represented in graph

8. The selected cluster which have the maximum score will be shown in thescatter plot as shaded.**[fig.5]**
9. The maximum test score computes the distance between each cluster.

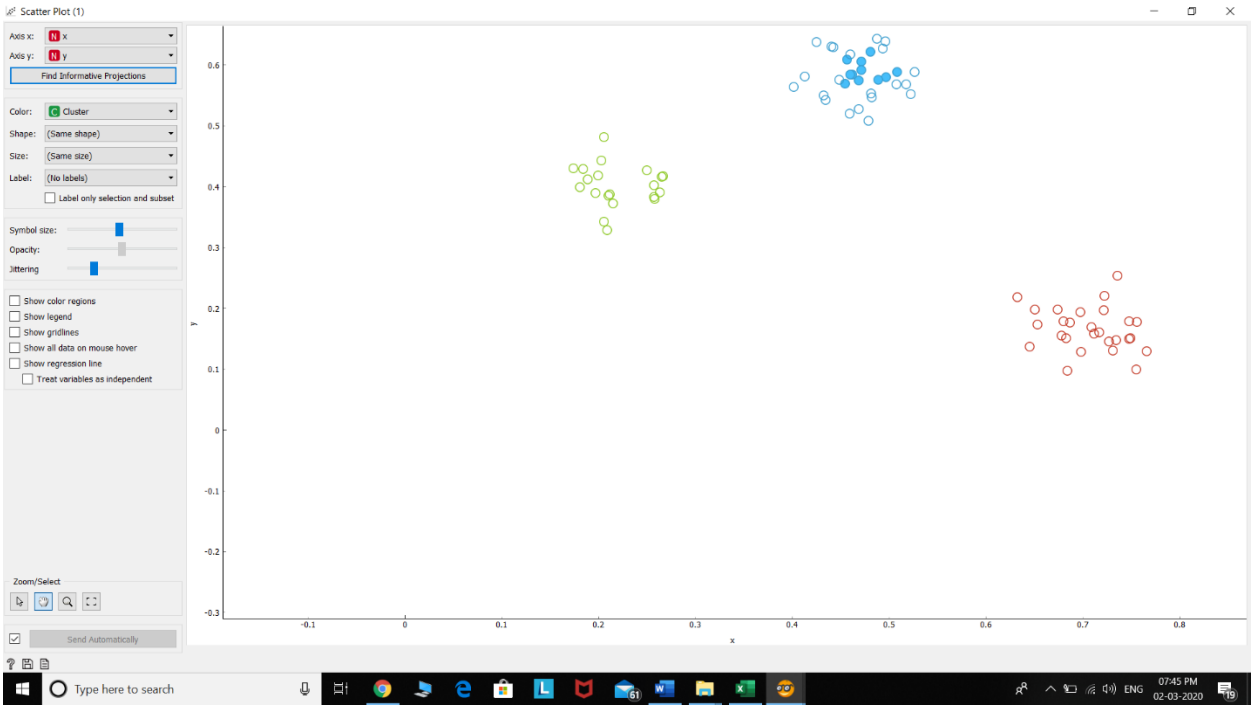


Fig.5 Workflow of scatter plot

CONCLUSION:

We have presented a dataset for clustering using Orange data mining, which we can understand easily by the graphical method. The proposed dataset consists of three clustering: Hierarchical clustering for extracting the data; K-means clustering for identifying the number of clusters: and Silhouette clustering for finding the distance of each other. Finally, we implemented data mining techniques which helpto identify the male and female affected by the skin cancer in different localization.

REFERENCES:

[1]. Gomathi S., and V. Narayani. "Systemic Lupus Erythematosus Prediction Tool UsingOptimal Cluster Based Classification (OCBC) Algorithm", International Journal of Engineering & Technology 7.4 (2018): 2806-2809

[2]. Gomathi, S., and V. Narayani. "Design And Implementation Of Optimal Chain Based Classification Algorithm (OCBCA) For The Early Prediction Of Systemic Lupus Erythematosus", International Journal of Pure and Applied Mathematics, 118,10 2018): 27-35

[3]. Gomathi, S., and V. Narayani. "Early Prediction Of Systemic Lupus Erythematosus Using Hybrid K-Means J48 Decision Tree Algorithm.", International Journal of Engineering & Technology 7.1.3 (2017): 28-32

[4]. Gomathi. S, Narayani. V, "Applying Decision tree algorithm to predict Lupus using Rapid Miner", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 9 (2015) pg 6728- 6731.

[5]. Gomathi. S, Narayani. V," Implementing Chi-Square Automatic Interaction Detection algorithm to predict Systemic Lupus Erythematosus", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.1 (2015) pp. 299-303

[6]. S. Gomathi, "A Deep learning of Autism Spectrum Disorder using Naïve Bayes, IBk and J48 classifiers", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-2, July 2019