International Journal of Computer Science Trends and Technology (IJCST) – Volume 11 Issue 5, Sep-Oct 2023

RESEARCH ARTICLE OPEN ACCESS

Hierarchical Clustering and Dendrogram Visualization

M. Gayathri [1], Sk. Johnbee [2]

Department of Computer Science, PB Siddhartha College of arts & science – Vijayawada

ABSTRACT

This paper represents a comprehensive study on data clustering analysis applied to live cosmetic product data obtained from the e-commerce platform Flipkart. The objective of this research is to uncover hidden patterns and groupings within the cosmetic product data, enabling insights for both consumers and businesses. The primary focus is on the generation of dendrogram diagrams that aid in visually understanding the inherent structure and relationships within the cosmetic product dataset.

Keywords: -Hierarchical clustering, Dendrogram, E-commerce

I. INTRODUCTION

Data mining involves the exploration and analysis of large and complex datasets to unearth valuable information that might not be immediately apparent. It draws upon various disciplines such as statistics, machine learning, and database management to uncover patterns, correlations and trends. By leveraging data mining techniques, organizations can gain a deeper understanding of their data, make predictions, and extract actionable insights. Data mining involves collecting raw data, preprocess the data for suitability for analysis, summarize and visualize data, discover hidden patterns, relationships and structures within the data, predict future trends, classify their attributes, group similar data points into clusters, items are discovered to uncover associations and dependencies.

A. S*upervised Learning*:

In supervised learning, the algorithm is trained on a labeled dataset, where the input data is paired with corresponding output labels. Supervised learning needs supervision to train the model, which is similar to as a student learns things in the presence of a teacher. Supervised learning can be used for two types of problems i.e. Classification and Regression.

B. Unsupervised Learning:

Unsupervised learning is another machine learning method in which patterns inferred from the unlabeled input data. The goal of unsupervised learning is to find the structure and patterns from the input data. Unsupervised learning does not need any supervision. Instead, it finds patterns from the data by its own. Unsupervised learning can be used for two types of problems: Clustering and Association.

Clustering:

Clustering is an unsupervised learning method, meaning it doesn't require pre-labeled data. It deals with the unlabeled data.

Clustering is a specific technique in data mining that focuses on grouping similar data points based on their similarities. The goal is to partition the data into subsets or clusters, where data points with in the same cluster are more similar to each other than to those in other clusters.

Agglomerative clustering:

Agglomerative clustering is a hierarchical clustering technique used in data analysis and machine learning. It is a bottom-up approach to clustering, where individual data points start as their own clusters and are progressively merged together to form larger clusters. This process continues until all all-data points belong to a single cluster

The exponential growth of online retail platforms like Flipkart has led to an enormous influx of data related to products, customer behavior and preferences. In this study, we harness the power of hierarchical clustering to uncover meaningful patterns with in the live cosmetic product data available on Flipkart. The utilization of dendrogram diagrams enhances our ability to comprehend the hierarchical relationships among cosmetic products.

II. METHODOLOGY

Initialization:

Begin by treating each data point as a single cluster, So, if you have 'n' data points, you start with 'n' clusters, each containing a single data point.

Calculate Pairwise Distances:

Compute the pairwise distances or similarities between all clusters by calculating Euclidean distance.

Merge Closest Clusters:

Based on the distance calculated in the above method merge them into a single cluster, reducing the total number of clusters by one.

Update Distance Matrix:

Recalculate the distances between the newly formed cluster and the remaining clusters. To do this compute the distance

between the new cluster and all other clusters using linkage criteria. Repeat the above steps to merging the closest clusters and updating the distance matrix until you have a single cluster that contains all data points. This forms a hierarchical structure represented as a dendrogram.

Dendrogram Construction:

The Dendrogram is a tree like diagram that illustrates the hierarchy of cluster merging. The vertical axis in the dendrogram represents the distances at which cluster were merged. The horizontal axis represents the clusters or data points.

III. IMPLEMENTATION

Data Collection:

We obtained live cosmetic product data from Flipkart through web scraping techniques. To do web scrapping we wrote a code in python to collect data attribute wise. We collect 2pages of data and it will be stored as a .csv file. We collect relevant attributes such as Product name, type, Brand, Original price, discount price, Rating. The following figure shows the .csv file in Microsoft excel format.

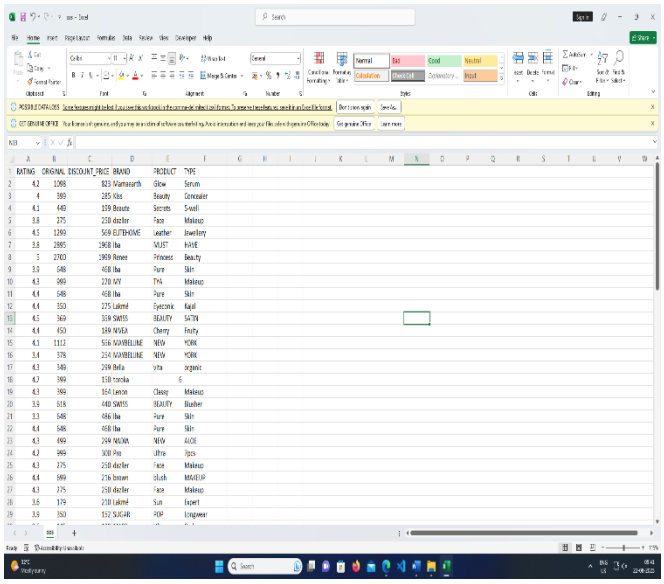

Fig. 1 Cosmetics data in .csv format

Data Preprocessing:

Preprocess the live data to prepare it for clustering. This may involve cleaning the data, handling missing values and converting textual information into numerical features using techniques like term frequency-inverse document frequency for product descriptions.

Data Selection:

Depending on the quantity of the data with its brand name, type we use selection method correlation analysis.

Hierarchical clustering:

Open weka and load preprocessed live data. Go to the "cluster" tab and select the "Hierarchical clusterer". Configure the clusterer settings, including the distance metric and linkage criterion, based on the nature of data.

But here to apply hierarchical clustering uses a Weka tool. So I converted the collected .csv to. arff file.

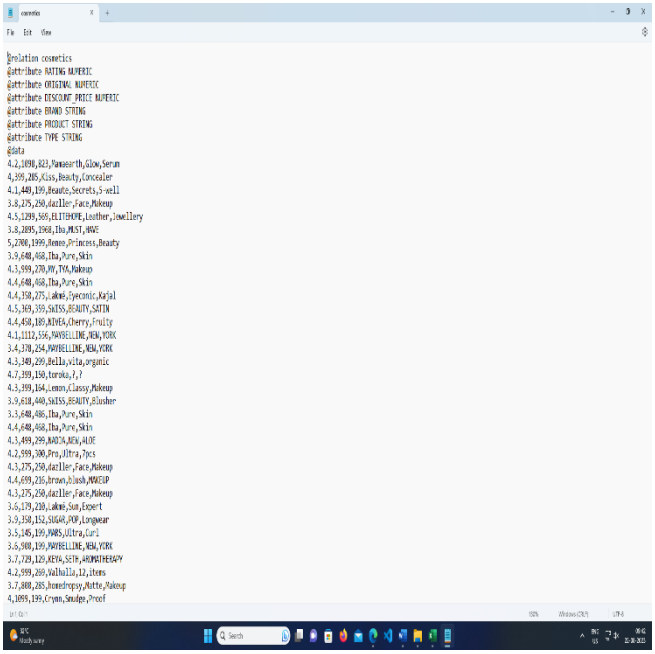

Fig.2 Cosmetics data in .arff format

Perform clustering:

Performed cluster hierarchical clustering on live cosmetic product data.

Cluster Analysis:

Based on the dendrogram and cluster assignments, analyze the clusters to identify patterns and similarities among cosmetic products.

IV. RESULT AND DISCUSSION

From the workflow built to obtain the result, the hierarchical clustering node

Then open the file in Weka tool and applied hierarchical cluster algorithm.

Then visualize the data in dendrogram.

Dendrogram Visualization:

Our approach generates dendrogram diagrams that reveal the hierarchical organization of cosmetic product based on their attributes, each node in the dendrogram represents a cluster, while the vertical axis depicts the dissimilarity between clusters.

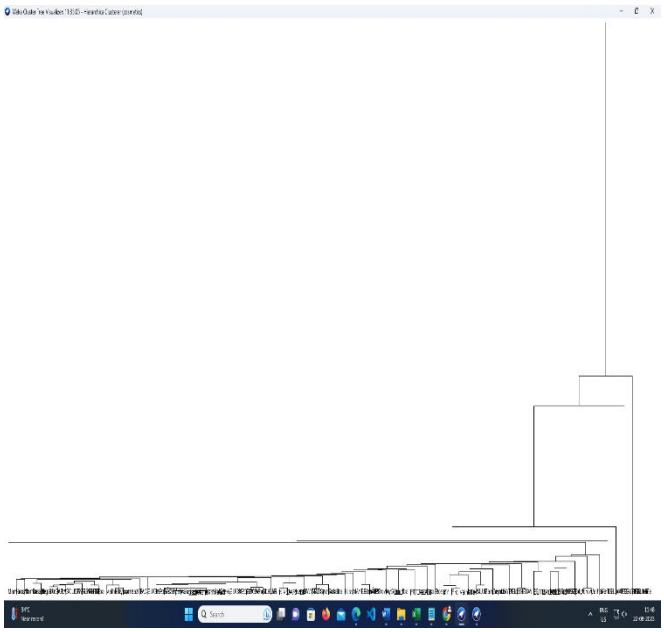

Fig. 3 Dendrogram diagram

Interpretation:

We can analyze the dendrogram to understand how the cosmetic products are hierarchically grouped based on their features and characteristics. We can explore different levels of the dendrogram to determine the number of clusters that make sense for our analysis.

Visualization and Reporting:

Finally based on the results of the hierarchical clustering we create visualizations or reports to stakeholders or for further analysis.

To keep your clustering results up to date

Implications and Applications:

The outcomes of this analysis hold significance for both consumers and businesses. Consumers can make informed purchasing decisions by understanding the grouping of cosmetic products that match their preferences. Businesses, on the other hand, can adapt marketing strategies, pricing tactics, and inventory management based on the identifies clusters.

V. CONCLUSION

This paper demonstrates the utilization of the weka tool for clustering analysis of live cosmetic product data from Flipkart. The dendrogram diagrams provide a visual representation of the underlying patterns and clusters with in the data. The findings contribute to a deeper understanding of consumer preferences and market dynamics in the cosmetic industry.

REFERENCES

1. Alkadhwi Ali Hussein Oleiwi, Adelaja oluwaseun Adebayo, Ali Alkattan Hussein, Data Mining Using Hierarchical Clustering Techniques on the position of employees in an information technology firm, Global Scientific Journals, GSJ: Volume 7, Issue 6, June 2019, Online: ISSN 2320-9186

2. Kaufman, L., &Rousseeuw, P.J. (2009). Finding groups in data: An introduction to cluster analysis. John Wiley&Sons.

3. Adelajaj O Adebayo and Mani S. Chaubey, “Data mining classification techniques on the analysis of student’s performance”. Global Scientific Journals (GSJ) ISSN 2320-9186, vol.7, issue 4, April 2019, pp 79-95

4. Shuhie A, Parul P and Seema M, “Hierarchical Clustering-An efficient technique of Data Mining for handling columinous data”. International Journal of Computer Applications (0975-887), Vol 129-No.13, November2015 pp 31-36

5. Szymkowiak A, Jan L, Lars K. H, “Hierarchical Clustering for Datamining”, pp. 1-5.

6. Fathi H S, Omer I E and Rafa E A, “Comparision of Hierarchical agglomerative algorithms for clustering medical documents”, International Journal of Software Engineering & Applications (IJSEA), Vol.3, NO.3, May 2012, pp1-15.

7. Odilie Y and Kylee. T. R., “Hierarchical Clustering Analysis: comparison of three linkage measures and application to psychological data”, The Quantitative Methods for Psychology (TQMP), Vol.11, no 1, 2015, pp 8-21.

8. Yogita. R and Dr. Harish R, “A study of hierarchical clustering algorithm”, International Journal of Information and Computation Technology.ISSN 0974-2239 © International Research Publications House. Vol. 3, No. 11(2013), pp. 1225-1232.