Data Mining Assignment - 2

Submitted by

Saikiran Challa

A Report submitted in partial fulfillment of the requirements for the Master of Data Science(Data Mining) Assignment

Data Mining Assignment - 2

Name : Saikiran Challa

Student No: 21356161

Degree: Master of Data Science

Lecturer: Lydia Cui

Lecturer: Lydia Cui,

Department of Computer Science and Information Technology

La Trobe University,

Bundoora, Victoria, 3086, Australia

Table of Contents

Task-1 : Data Preprocessing

Introduction:

Data preparation is an important stage in clustering because it assures equal contribution from all characteristics and prepares the data for effective segmentation. Given the nature of clustering algorithms, adequate management of categorical variables and standardisation of numerical data are required.

Missing Values:

The dataset was examined for missing values, and no missing values were found. This guaranteed that all data entries were complete and trustworthy for clustering.

One-Hot Encoding:

Categorical variables, notably 'Channel' and 'Region,' were one-hot encoded to yield numerical representations. This approach prevents the introduction of unintentional biases and enables clustering algorithms to analyse these variables properly.

Feature Scaling:

The numerical characteristics ('Fresh', 'Milk', 'Grocery', 'Frozen', and 'Delicassen') were standardised with StandardScaler. Standardisation was required because distance-based clustering methods, such as K-Means and hierarchical clustering, are sensitive to the scales of the input characteristics. Without scaling, characteristics with higher ranges may dominate the clustering process.

Data Subsampling:

A subset of the data (10%) was utilised for first procedures to improve performance and conduct preliminary analysis. However, the entire dataset was used for the primary clustering tasks to provide reliable findings.

Task-2 : Implement K-Means and Hierarchical Clustering

K-Means Clustering:

- The elbow approach was used to calculate the ideal number of clusters (k). The sum of squared errors (SSE) was displayed against various values of k ranging from 1 to 10. Based on the elbow plot, k = 4 was selected as the ideal number of clusters since it reflected the point at which adding more clusters did not significantly lower the SSE.

- Justification: "The choice of k = 4 was supported by looking at the elbow point in the figure, which shows the declining rewards of adding more clusters. The elbow point indicates a compromise between underfitting (too few clusters) and overfitting (too many clusters).

- The visualisation follows: "A 2D scatter plot was created to visualise the clustering results, with each cluster represented by a different colour, and the centroids marked for easy interpretation."

Hierarchical Clustering:

- **Dendrogram:** "Hierarchical clustering was done using the 'ward' linking approach, and a dendrogram was generated. The dendrogram helps visualise the merging process, allowing for the determination of an appropriate number of clusters. The results were consistent with the K-Means clustering, showing that k = 4 was an appropriate option."

- **Interpretation:** "The hierarchical dendrogram revealed insights into the dataset's structure, demonstrating how groups join at various distances. This allows for a better grasp of the linkages between data pieces.

Task-3 : Compare and Analyze Clustering Results

Cluster Centroids (K-Means):

- "The clusters' centroids reveal the general features of each segment. For example, Cluster 1 had greater values in 'Milk' and 'Grocery,' indicating consumers who spent a lot of money in these categories, potentially indicating wholesale clients."

- **Explanation:** "By evaluating each feature's centroid value, we may gain insight into spending trends across consumer segments. Cluster centroids assist in identifying common characteristics among customers, such as preferences for specific product categories."

- **The Table's Format:** "A table summarising the centroid values for each feature in the four clusters is provided, highlighting key differences between the clusters."

Silhouette Scores:

"Silhouette scores were computed for K-Means (0.3546) and Hierarchical Clustering (0.2451). The ratings show how well-defined the clusters are, with larger values indicating stronger separation. The higher score for K-Means suggests that it produced more identifiable and cohesive groups than hierarchical clustering."

Strengths and Weaknesses:

- **K-Means:** "K-Means is computationally efficient and ideal for spherical clusters, but it is sensitive to outliers and requires k to be specified in advance. It works best with datasets that have clearly defined, similarly sized clusters."

- **Hierarchical Clustering:** "Hierarchical clustering does not require a predetermined number of clusters and can reveal data structure, but it is computationally expensive and less effective for large datasets." It is more adaptable to different cluster sizes and forms."

Task-4 : Discussions

Comparison of Unsupervised (K-Means) vs. Supervised (SVM) Methods:

- "K-Means clustering is an unsupervised approach that divides clients based on intrinsic patterns in the data rather than using pre-defined labels. In contrast, SVM (a supervised approach) was employed to predict 'Region' from input characteristics."

- **SVM Performance:** "The SVM classification report demonstrated an overall accuracy of 84%, although it performed badly in predicting 'Region 2' due to unbalanced data or overlapping feature spaces. The report includes information on each class's precision, recall, and F1 scores.

Business Strategies Based on Clusters:

- Customised Marketing: "K-Means clusters may be utilised to design marketing strategies to specific client categories. For example, a cluster with high spending on 'Fresh' items may benefit from bulk buy deals, whilst those with strong spending on 'Grocery' could be targeted with discount bundles.

- **Regional specials:** "Using SVM to identify regions allows businesses to better understand regional preferences and offer localised specials or region-specific items. For example, if 'Region 3' has a large concentration of clients from Cluster 2, promotions can be targeted to this segment's purchasing patterns."

- **Loyalty Administration:** "Insights from clusters can also help develop loyalty programs that encourage repeat purchases by rewarding specific spending behaviours identified in key customer segments."