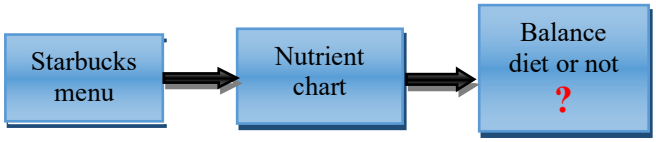
Categorizing unlabeled data
(A case study based on Starbucks)

I. Motivation

A. Problem

Starbucks is the world's largest coffeehouse chain. Starbuck aim to ensure that everything they sell is produced to high quality and ethical standards. They have great menu consist of varieties of food and drinks. Researchers always suggest eating the fast-food item in moderation due to high calories, sugar and sodium. Though Starbucks have a good nutrient chart but those are not specified whether its is healthy or not for specific person. In other word these foods are not arranged or grouped based on nutrient values.

The coffeehouse website is very rich in information and well-designed which focus on nutritious value. Moreover, people are becoming health conscious and are more concerned with nutrition fitness. For these people the regular menu may not contain any specific choice or there is not specific label that gives the idea whether the food is suitable for diabetic or hypercholesterolemia patient. Now no one will count all the nutrient before ordering a food. Also, the data they have does not give any specific pattern or idea about specific food. We want to find out the group foods that unhealthy. This may help to group certain food that are healthy and can include them in healthy eating menu.



B. Kind of problem

The nutrient chart of Starbuck contains nutrient values but are not categorized. The data contain information but using that information we want to recognize some pattern. This may help the coffee shop to separate lower calories food or to lower some element from food to make the unhealthy food healthier.

The above-mentioned problem can be categorized as clustering problem. The data set have different nutrient percentage in specific food but those are not labeled or tagged.

Labeled data are the data that contain an identifier or specific tag. On the other hand, unlabeled data are without tag and difficult to classify or categorize. The nutrient chart of Starbuck contains nutrient value but are not classified according to calories.

C. Necessity to solve the problem

In supervised learning, the data carry specific identifier or label but in unsupervised learning data contain different features but no specific label. Clustering is the task of grouping similar unlabeled data together. On the other hand, classification is diving labeled data into different classes.

In the above-mentioned problem (Starbucks) the data set contain different features like name, calories, fats, sodium, cholesterol but the food item does not contain any specific label like low sugar or low cholesterol.

We want to group data based on similar characteristics and find out whether clustering help us or not to categorize the data into different classes. Grouping item will help to find food that are unhealthy with more than one nutrient values. In addition, they can offer some special menu like low sugar or low-calorie, low cholesterol items.

D. Difficulties in solving the problem

Clustering is the process of grouping item of similar characteristics together. On the other hand, classification is the task of dividing data into different classes or categories. The data we are working with is unlabeled data. Unlabeled data or data with very prior domain knowledge always struggle with computational cost.

For solving the problem mentioned above we have first applied clustering algorithm and based on the result we have then prepared a classification model.

For clustering we are using hierarchical clustering and k-means clustering on different features (calories, fat, sugar, cholesterol). Due to n iteration and repeated update in matrix, the time complexity of agglomerative hierarchical clustering is very high. For large dataset the algorithm does not perform very well.

However, we have used K-means algorithm to eliminate time complexity. But in k-means choosing number of clusters is sometimes a struggle. The main disadvantage of k-means is its accuracy. Sometimes due to manual selection of cluster numbers, some data point may form wrong group.

Moreover, we are using k-nearest neighbors for classification. Based on the clustering result we want to categorize the data in various classes (using clusters as label) so that it can be used later for prediction or improvement in supervised learning.

As the dataset is unlabeled and have various features so time complexity is an issue to solve the problem.

	Hierarchical clustering	K-means clustering	k-nearest neighbors
Time complexity	$O(n^3)$	O(t*k*n*d)	Training-O(d) Testing-O(nd)

II. State of the Art

Machine learning is the branch of computer science and artificial intelligence that focus on data and can improve automatically through experience. In other word it teaches machines how to use data efficiently and extract information from the data that we cannot sort [1]. Machine learning relies on different techniques and algorithm to extract information based on the type of data. Supervised machine learning aims to produce general pattern and predict future instances [2]. On the other hand, in unsupervised learning, algorithm learn pattern from untagged data. There is a similarity between unsupervised learning and statistical modeling. Both provide outputs without supervision [3]. Clustering is a part of unsupervised learing which discover groups in unlabeled data. Clustering analyses are used in various field like disease analysis and finding similar characteristis in in diseases [4].

Several approaches of clustering are developed for grouping data. Among different clustering technique Hierarchical clustering present the clusters as tree structure. Agglomerative Hierarchical Clustering (bottom-up) work by dividing all data point to single cluster and then join them together to form a single cluster based on distance. Agglomerative clustering is widely used to recognize hidden pattern in unlabeled or unstructured data. It is sometimes used to find the frequency of similar item that has been grouped together [5]. The purpose of clustering is separate a set of N object into C cluster so that similar object remains in one group and dissimilar in another group. In Agglomerative clustering, split and merge are determined in greedy manner. Generally, the output is represented by Dendrogram [6].Usually, in Agglomerative Hierarchical clustering(AHC) after individual cluster formation a distance matrix is genarated for dataset X with maximum matrix size (n×n) based on different distance matrics like Euclidean Distance, Manhattan Distance, Minkowski Distance, Hamming Distance.The distance between two clusters are calculated besed on various linkage metrics.Among them ward method is used to reduce the sum of squared error among the clusters.The computational complexity of AHC is o^3 for (n-1) iterations(where n is the number od data object in dataset) [7].

K-means algorithm is well-known for its efficiency with large unlabeled dataset. K-means algorithm mainly aims to select central point (centroid) and calculating Euclidean Distance. It find the center of cluster (centroid c1,c2…c_k) in a way that sum of squared distance of each datapoint xi, $1 \leq i \leq n$ to its closest center point is minimized. Sometimes enhancing Euclidean distance formula can increase the accuracy and prediction level of k-means [8]. We can determine the number of clusters using various method like Elbow method, silhouette method, Gap static method. Silhouette score is calculated based on silhouette coefficient. Silhouette score S(i) determines how well each clustered are formed with similar object.

$$S(i) = \frac{b_i - a_i}{\max\{b_i, a_i\}} \qquad \text{Where, } -1 \leq S(i) \leq 1$$

In k-means choosing the right number of clusters is very crucial. Changing the number of clusters change the average score of silhouette score which help us to determine the best cluster [9]. One issue with k-mean clustering is visualization with multidimensional data. Usually, the cluster can be represented in 2d or 3d but for more three-dimensional data we have to rely on mathematical calculation rather than visualization.

K-nearest neighbors algorithm is know as an effective lazy classifier. It classifies any data by searching for the nearest neighbor and combine the together to form one cluster. The parameter k decides how many neighbors should be considered for observation. The algorithm depends on past observation for future prediction. The algorithm is highly sensitive to noisy data and missing data.

Data reduction technique can improve the computational cost [11].

Classification via clustering can be used to predict future data. Accuracy between Hierarchical Clustering algorithm and k-means algorithm which support classification via clustering, the accuracy of k-means seems better. For this purpose, clustering is first done with different clustering algorithm then the traditional classification algorithm is applied to check the accuracy. Accuracy between Hierarchical Clustering algorithm and k-means algorithm which support classification via clustering, the accuracy of k-means seems better. [12]. In one experiment it has been observed that the accuracy of k-mean algorithm increases with the increased number of clusters while performing classification using unlabeled data [13].

k-means is an unsupervised learning algorithm whereas k-nearest neighbors algorithm is an supervised learning algorithm. In k-means the value k decide the number of clusters and in k-nearest neighbors the value of k decide the number of neighbor to be considered for observation. In k-nearest neighbors algorithm we need to preprocess the data set but for k-mean it is not necessary to do scaling. K-nearest neighbor algorithm be used as classifier or regression. In classifier it represent the accuracy of test vs training sample.

In many cases these two algorithms have been combinedly used for text recognition. After clustering the cluster centers are defined as category. In this way the accuracy increases and time for calculating similarities decreases [14]

In another study for text classification, k-nearest neighbors and k-means algorithm has been used together. The simulation result shows that using k-mean and k-nearest neighbors together reduced the number of training sample and computational complexity [15].Clustering itself may not be enough to classify data but it can be used as a feature to improve the accuracy of classification.

III. Solutions

For finding pattern or group in a dataset we can apply clustering algorithm. To find the structure from the unstructured data we first use Hierarchical clustering based on two dimensions (Cholesterol[mg] and Sodium[mg]). Then we have applied k-mean clustering featuring sugar and

calories. Then after finding group within the dataset, we used the clusters as label and applied k-nearest neighbors to divide the data in different classes to predict future data.

A. Hierarchical clustering (AHC)

Algorithm

Input- dataset df

Output-A tree structure cluster assignment

Start

Step 1-importing food menu dataset by pandas

Step 2-Normalize DataFrame

Step 3-Create Dendrogram

X-axis contain number of food item, Y-axis contain Euclidean distance

Step 4-Determine the optimal number of Clusters based on dendrogram

Step-5-Apply Agglomerative Clustering

(n_clusters=2, affinity='euclidean',linkage='ward')

Euclidean distance

$$d(p,q) = \sqrt{\sum_{i=0}^{n}(q_i - p_i)^2}$$

Step-6 Fit hierarchical clustering algorithm to dataset and predict food items that belongs to same group

cluster.fit_predict(df)

step-7 Plot the clusters

step-8 Visualize the result

End

B. K-means clustering

Algorithm

Input- dataset with x data points

Output-k no of cluster assignment

Start

Step 1-import 'Starbucks.csv' dataset

Step 2-Normalize DataFrame

Step 3- Plot dimensions [Calories],[Sugars(g)],[Total Carbohydrate(g)]

Step 4-Determine number of Clusters(k).

Step-5-Determine the centroid c_1,c_2…c_k randomly

km=KMeans(n_clusters=2,random_state=1)

Step-6 For each datapoint X_i, find closest centroid and assign data points to that cluster(k)

step-7 Compute and place new the centroid to each cluster

step-8 Repeat until no reassignment occurs.

step-9 Visualize the Result (Clusters)

End

C. k-nearest neighbors

Algorithm

Input- dataset with x data points

Output-Classification model

Start

Step 1-Import 'Starbucks.csv' dataset

Step 2-Split dataset to attribute and labels

Step 3- Train and test split

Step 4- Determine the value of k

Step-5- For each data, Calculate Euclidean distance between test and train

Based on result sort them in ascending order

Step-6-Assign class to the test point based on the most frequent class of these row

step-9 Return model evaluation

End

IV. EXPERIMENT AND RESULT

A. Experiment with Hierarchical clustering

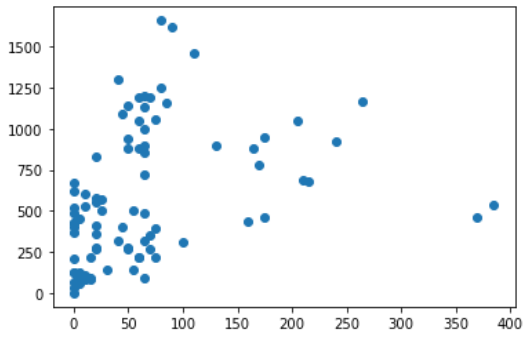


Fig- Scatter plot of dataset df before clustering

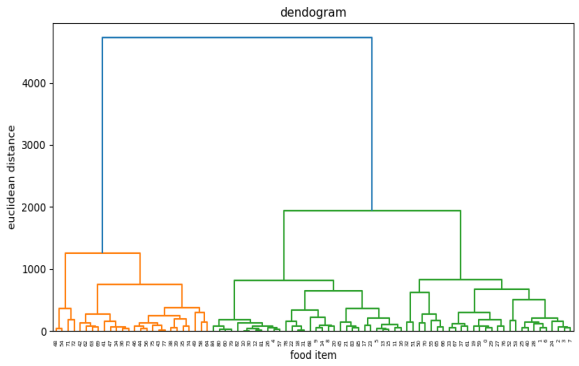


Fig- Dendrogram for optimal clustering

	Name	Calories	Sugars(g)	cluster
0	Chonga Bagel	300	5	0
1	8-Grain Roll	340	15	0
2	Almond Croissant	420	13	1
3	Banana Nut Bread	420	30	1
4	Birthday Cake Pop	170	18	0
..	...	...	...	...
81	Justin's Chocolate Hazelnut Butter	180	7	0
82	Justin's Classic Almond Butter	190	1	0
83	Lemon Crunch Yogurt Parfait	330	29	0
84	Mango & Coconut Yogurt Bowl	250	26	0

Fig- clusters are assigned to different food items

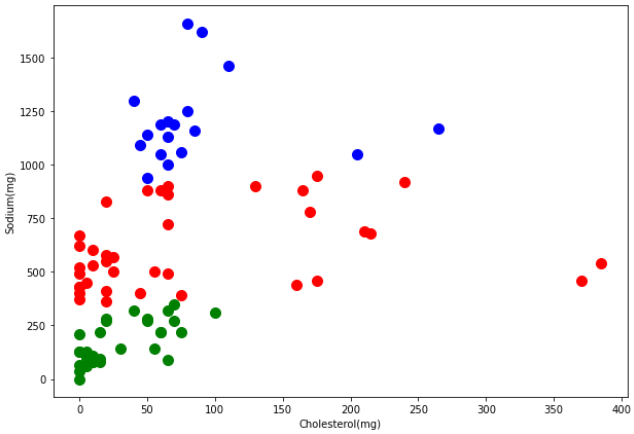


Fig- Visualization of result with cluster =3

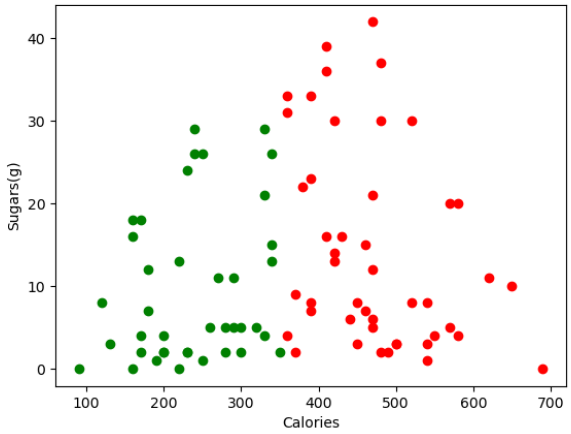


Fig- visualization of result with cluster =2

B. Experiment with k-means

C. Experiment with k-nearest neighbors

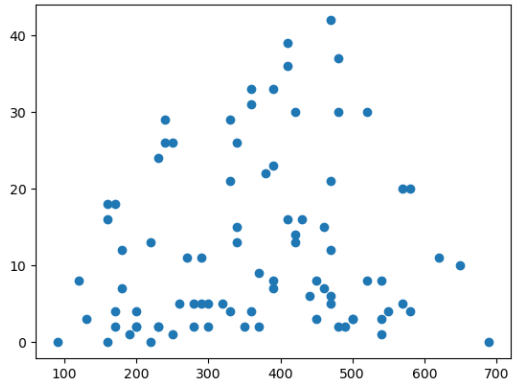


Fig- Scatter plot of data point (sugar, calories, carbohydrate)

	precision	recall	f1-score	support
0	0.79	0.92	0.85	12
1	0.95	0.87	0.91	23
accuracy			0.89	35
macro avg	0.87	0.89	0.88	35
weighted avg	0.90	0.89	0.89	35

Fig- Result of evaluation model

D. Result

Agglomerative Hierarchical clustering

Cluster	Silhouette score
Cluster=2	0.5445
Cluster= 3	0.4646

k-mean silhouette score

Cluster (k)	Silhouette score
K=2	0.5714
K=3	0.5331
K=4	0.5459

Silhouette score vs Accuracy

Silhouette score of clustering	Accuracy of evaluation model
K=3, S=0.5331	Accuracy = 0.83
K=2, S=0.5714	Accuracy = 0.89

Execution time

Algorithm	Time taken to execute
Agglomerative clustering	3.8190436363220215
k-means	1.565791130065918
k-nearest neighbors	1.026252269744873

V. SUMMARY AND OUTLOOK

Data is the ground based on what machine learning adapt information and build knowledge. The main concept of machine learning algorithm is once it learns the pattern then it can predict future. We have applied the same solution to solve our problem.

The data set [10] we have used does not contain any tag. For categorizing the data into different category we have used two clustering algorithms. The data contain information about various nutrient value in different food item. Our aim is to learn the hidden pattern and group the data into classes so that we can identify which food items are healthy and which food items are unhealthy.

At first, we have applied agglomerative hierarchical clustering algorithm in for dataset X with column [Cholesterol(mg), Sodium(mg), Total Fat(g)]. The result shows 3 cluster. The green cluster shows food item with lowest amount of cholesterol, fat, and sodium. Red cluster represent medium level of sodium with medium to highest level cholesterol. The blue cluster represent highest level of sodium with moderate level of cholesterol. However, the overall result of hierarchical clustering was not satisfactory and the time complexity of algorithm is very high.

The second approach we have used is k-means clustering algorithm. K-means is a centroid based algorithm which calculate the distance of each data point from centroid and cluster the nearest one. In k-means algorithm the importance of k is vital. The variable K is a hyper parameter whose value we set before training. The value of k is the number of clusters we want. For getting the optimal number of clusters we have calculated the silhouette score for each cluster, and we have decided to cluster them in two groups. The result shows that green cluster have lowest calories and low amount of sugar. The items have been listed as cluster 0. On the other hand, we have red clusters which shows high calories of food with good amount of sugar and carbohydrate. Time complexity of k-mean is comparatively low and the silhouette score with cluster =2 is highest of all. Though silhouette score is used to find the goodness of clustering but finding the accuracy of clustering algorithm is difficult as there is no supervision present.

In unsupervised learning clustering is used to label the data and classify the data into classes. We have used the cluster as a label and applied k-nearest neighbors to train a classification model. The accuracy of the model seems 89%. The time complexity of k-nearest neighbors is comparatively low then k-mean. Also, the accuracy rate seems to improve when the silhouette score rises.

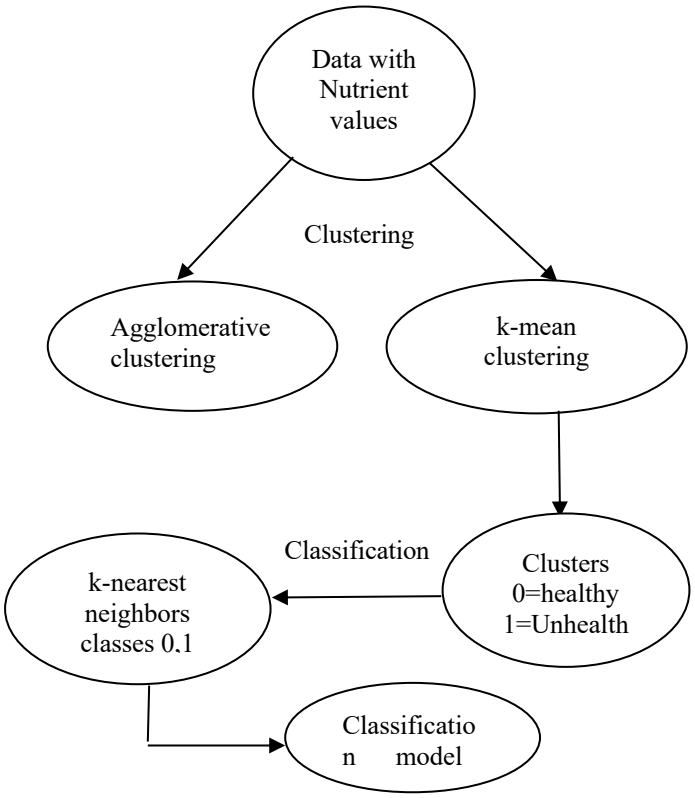


Fig-Solution Approaches

Hierarchical clustering work better with low dimension data and it is good for checking the similarities or relationship between sub clusters. K-mean clustering work perfectly with large dataset but choosing the right value of k is the main drawback of this algorithm. Another tendency of k-mean algorithm is to build equal size clusters. But for unlabelled

dataset and pattern recognition k-means algorithm work best. Moreover, if the dataset is labelled then classification is an easy way to build a classification or evaluation model. But we can use unlabelled data to improve supervised learning or get an idea about how well the classes are formed. Clustering itself may not be enough to classify data but it can be used as a feature to improve the accuracy of classification.

REFERENCES

[1] B. Mahesh, "Machine Learning Algorithms - A Review," International Journal of Science and Research (IJSR), vol. 9, no. 1, Jan. 2020.

[2] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 1310-1315.

[3] Z. Ghahramani, "Unsupervised learning guide [Zoubin Ghahramani]," Unsupervised Learning∗, 16-Sep-2004. [Online]. Available: https://datajobs.com/data-science-repo/Unsupervised-Learning-Guide-%5bZoubin-Ghahramani%5d.pdf. [Accessed: 25-Jan-2022].

[4] K. XIA, Y. WU, X. REN, and Y. JIN, "Research in clustering algorithm for diseases analysis," Journal of Networks, vol. 8, no. 7, 2013.

[5] R. W. Sembiring, J. M. Zain, and A. Embong, "A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course," JOURNAL OF COMPUTING, vol. 2, no. 12, Dec. 2010.

[6] k Sasirekha and P. Baby, "Agglomerative Hierarchical Clustering Algorithm- A Review," International Journal of Scientific and Research Publications, vol. 3, no. 3, Mar. 2013.

[7] S. K. Seetharaman, M. Muthusamy, M. S. H, and V. B. A, "A brief survey of unsupervised agglomerative hierarchical clustering schemes," International Journal of Engineering & Technology, vol. 8, no. 1, 2019.

[8] A. Bansal, M. Sharma, and S. Goel, "Improved K-mean clustering algorithm for prediction analysis using classification technique in Data Mining," International Journal of Computer Applications, vol. 157, no. 6, pp. 35–40, 2017.

[9] H. W. Choi, N. Muhammad Faseeh Qureshi, and D. R. Shin, "Comparative analysis of electricity consumption at home through a silhouette-score prospective," 2019 21st International Conference on Advanced Communication Technology (ICACT), 2019.

[10] "Starbucks nutrition (with sugar and etc..)," Kaggle, 10-Oct-2019. [Online]. Available: https://www.kaggle.com/swoolfeek/starbucks-nutrition-with-sugar-and-etc. [Accessed: 31-Jan-2022].

[11] S. Ougiaroglou and G. Evangelidis, "Fast and accurate K-nearest neighbor classification using prototype selection by clustering," 2012 16th Panhellenic Conference on Informatics, 2012.

[12] Lopez, M.I., Luna, J.M., Romero, C. and Ventura, S., "Classification via clustering for predicting final marks ...," Classification via clustering for predicting final marks based on student participation in forums,2012.[Online].Available: https://files.eric.ed.gov/fulltext/ED537221.pdf. [Accessed: 31-Jan-2022].

[13] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms," Proceedings of the 2006 SIGCOMM workshop on Mining network data - MineNet '06, 2006.

[14] P. WiraBuana, S. Jannet D.R.M., and I. Ketut Gede Darma Putra, "Combination of K-nearest neighbor and k-means based on term re-weighting for classify Indonesian news," International Journal of Computer Applications, vol. 50, no. 11, pp. 37–42, 2012.

[15] Y. Zhou, Y. Li, and S. Xia, "An improved KNN text classification algorithm based on clustering," Journal of Computers, vol. 4, no. 3, 2009.