Linguistic Summarization of Clusters

Abhishek Pande (17115005), B. Tech III yr.
Arpit Gupta (17116015), B. Tech III yr.
Electronics and Communication Engineering

There is a rapid rise in the amount of data generated in today's digitalized world. As a result, it is increasingly hard to gain useful information from the vast amount of data produced. Data mining comes handy as it helps in revealing unsuspected patterns and trends in the unstructured data. Such patterns can then be seen as a kind of summary of the input data and may be used in further analysis. Linguistic summarization helps to point analysts in the direction of useful information, by verbalizing interesting hidden patterns in the data. In this paper, we aim to extract useful insights after clustering a given dataset into clusters. The proposed paper provides a way for cluster analysis/evaluation and works for various kinds of datasets such as Iris and, Wheat seed.

Introduction: Clustering is the technique of grouping data based on some similarity. Defining what we mean by "similar" observations, is a crucial aspect of cluster analysis, which usually requires a lot of contextual expertise and creativity beyond what statistical tools can present. This analysis is achieved by minimizing intragroup similarity and maximizing inter-group similarity; in other words, members of a given cluster are more like each other than those of other cluster members. Clustering is an unsupervised machine learning process that is fundamental to data mining. Many data mining queries deal either with how the data points are organized or which points could be considered remote from natural groupings.

There exist a large number of clustering algorithms. These clustering algorithms can club into: hierarchical methods, partitioning methods, grid-based methods, and density-based methods. In order to examine the results of experimentation, we used k-Means and DBSCAN as clustering algorithms. The K-Means [1] approach determines k group representatives and casts each data point to the cluster with its closest representative to the object. The sum of the square of distances between the data points and their representatives is minimized. DBSCAN [2] locate dense regions that are parted by low-density regions and groups together the data point in the same dense region.

In an insight discovery process, analysis and evaluation of the results are fundamental in practice. In the data clustering case, however, it is usually difficult to observe in what condition each cluster forms. Visualization is a common way of grasping things. Nevertheless, there would be a high implementation cost or a physical constraint in visualization. So, we would like to verbalize the clusters [4], i.e., ally a perceptive descriptive summary with each cluster. Also, it seems desirable that the summary is objective and is automatically chose from clusters.

K-Means Clustering: It is an iterative algorithm that tends to divide the dataset into *k* specific non-overlapping clusters where each object belongs to a specific cluster. It uses "Expectation Maximization" to solve the clustering problem. The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

Since it uses distance-based measurement to find the similarity between the objects, it's recommended to normalize the data to have zero means and a standard deviation as one. Sometimes this algorithm may not converge to the global optimum and may be stuck into the local one. So, it should be run using different centroid initialization to yield a minimum sum of squared distance.

DBSCAN: DBSCAN is a density-based clustering algorithm that learns to separate groups of high density from low-density groups. The algorithm forms an *n-dimensional* shape around the object and estimates how many points fall within that shape. It then iteratively develops the cluster by passing through each data point. It can also handle the outliers present in the low-density region. But it struggles with identifying clusters in data sets that contain nested or overlapping clusters of varying or similar densities. It requires two parameters:

eps: specifies how close data points should be to each other to be treated as a part of a cluster.

minPoints: the minimum number of data points to structure a dense region.

Performance Measure: There are mainly two types of measures to evaluate the clustering performance. The first one requires a ground-truth label and is known as Extrinsic Measures. The later one does not need the ground truth labels and is called Intrinsic Measures. For our experimentation, we have used the following types:

1) Silhouette Coefficient: An example of an intrinsic measure where evaluation performance takes place using the model itself. It ranges from [-1,1]. A higher Silhouette Coefficient [3] score means better-defined clusters in the model. The zero scores indicate overlapping clusters. The score is on a higher side when clusters are well separated and dense. The Silhouette Coefficient is assigned for each data sample and consists of two scores:

a: The mean distance between a data sample and all other data points in a given cluster.

b: The mean distance between a data sample and all other data points in the next closest cluster.

Thus, Silhouette Coefficient 's' for a sample is:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

2) Homogeneity, completeness, and V-measure score: If we have knowledge about the ground truths of the data, we can possibly define some intuitive measure. For any cluster assignment, we have the following two desired objectives:

homogeneity: defines that each cluster confines only samples of a single class.

completeness: defines that all the samples of a given class are assigned to the same cluster.

These concepts can be modeled as scores (homogeneity score and completeness score) for evaluating the clustering. They are bounded by [0,1]. Moreover, the harmonic mean of them gives a V-measure score.

Experimentation: In this part, we show some of the experimental results. Since the class information won't be available in unsupervised learning, we first estimate the number of clusters using the Elbow method as a starting point. Also, some data preprocessing techniques like Min-Max Scaler or standardization are employed to bring the data features in a range of [0,1] or [-1,1]. For multi-dimensional data, PCA (principal component analysis) can be used to reduce its dimensions to a particular number. PCA is an orthogonal linear transformation that converts the data points to a new coordinate system such that the largest variance by some scalar projection of the data lies on the first principal component (first coordinate), the second greatest variance on the second coordinate, and so on. The code for our experimentation is available here.

General Summarization: We defined a few parameters on which we created a general summarization of any dataset:

- Cluster center distance.
- Maximum and Minimum data point for every individual cluster.
- Variance of all the data points that belong to a cluster.
- Labels of the predicted cluster.
- Silhouette score or different performance measures.
- Total number of outliers of the cluster. Outliers are defined as the objects that lie outside of the 90% of the maximum distant data point radius of a cluster.

The procedure followed to generate the summary:

1. Get the optimum number of clusters for a dataset. We used the elbow method for this. The **Elbow method** runs K-Means clustering on the dataset for different values of the number of clusters. It then calculates the average score based on different metrics like the Silhouette score and plots it as score vs. clusters. The elbow that is the inflection point is the best value for k.
2. Apply the K-Means algorithm to get the labels for each data point of the data set.
3. Separate different cluster data points by mapping them to their corresponding labels. Then average cluster center is generated by taking the average distance of all the data points of their corresponding labels to get the closest K-Means generated center for their corresponding cluster.
4. Apply different metrics to calculate the maximum, minimum

distances, the variance, and all the other summarization points, as mentioned above.

Datasets used to generate their Linguistic Summary:

a). Iris Dataset: This dataset contains a set of 150 data points under five attributes - Petal Length, Petal Width, Sepal Length, Sepal Width, and Label (Species). It contains three labels/classes of 50 instances each, where each label refers to a type of iris plant. These features are measured in centimeters. All three species in the data are separable in the projection on the branching principal and component nonlinear.

The Elbow method is used to find the value of *k*.

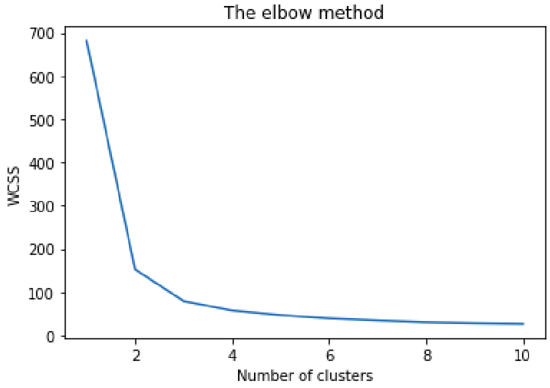
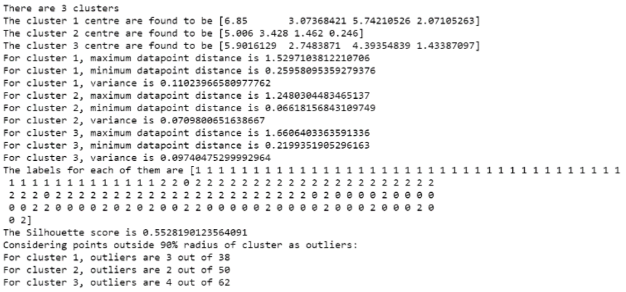


Fig 1. Elbow method for Iris dataset

So, from the above curve, we get a number of clusters as 3.
The generated summary:



```
There are 3 clusters
The cluster 1 centre are found to be [6.85       3.07368421 5.74210526 2.07105263]
The cluster 2 centre are found to be [5.006 3.428 1.462 0.246]
The cluster 3 centre are found to be [5.9016129  2.7483871  4.39354839 1.43387097]
For cluster 1, maximum datapoint distance is 1.5297103812210706
For cluster 1, minimum datapoint distance is 0.25958095359279376
For cluster 1, variance is 0.11023966580977762
For cluster 2, maximum datapoint distance is 1.2480304483465137
For cluster 2, minimum datapoint distance is 0.06618156843109749
For cluster 2, variance is 0.0709800651638667
For cluster 3, maximum datapoint distance is 1.6606403363591336
For cluster 3, minimum datapoint distance is 0.2199351905296163
For cluster 3, variance is 0.09740475299992964
The labels for each of them are [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0 2 0 0 0 0 2 0 0 0 0
 0 0 2 2 0 0 0 0 2 0 2 0 2 0 0 2 2 0 0 0 0 0 2 0 0 0 0 2 0 0 0 2 0 0 0 2 0
 0 2]
The Silhouette score is 0.5528190123564091
Considering points outside 90% radius of cluster as outliers:
For cluster 1, outliers are 3 out of 38
For cluster 2, outliers are 2 out of 50
For cluster 3, outliers are 4 out of 62
```

b). Wheat Seed Dataset: It contains 70 elements each of three different varieties of wheat seeds: Kama, Rosa, and Canadian. The seven seed variables are Area, Perimeter, Compactness, Kernel Length, Kernel Width, Asymmetry Coefficient, and Kernel Groove Length. The last column in the data is reserved for the Kernel type. This particular dataset has 199 entries. Some of these variables are explicitly dependent. For e.g., compactness: C = 4*π*Area/(Perimeter)^2 has a linear proportional relationship with the area, and also a square proportionality with kernel width.

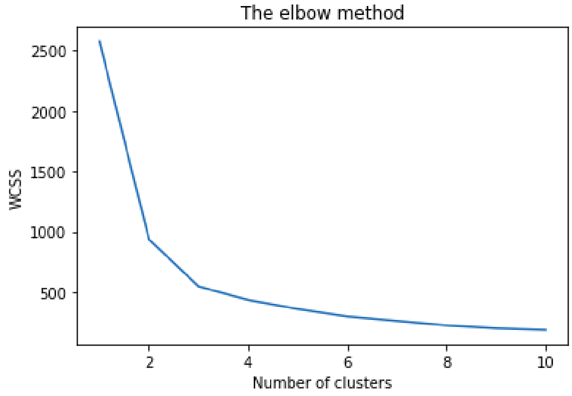
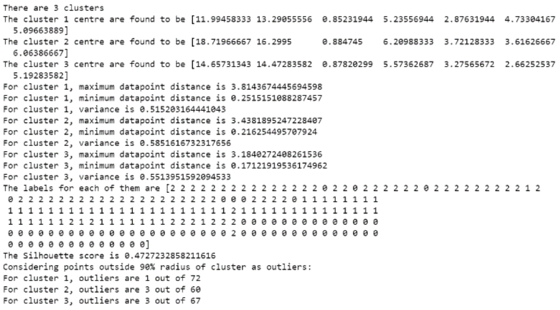


Fig 2. Elbow method for Wheat seed dataset

So, from the above curve, we get a number of clusters as 3.
The generated summary for this dataset:



```
There are 3 clusters
The cluster 1 centre are found to be [11.99458333 13.29055556  0.85231944  5.23556944  2.87631944  4.73304167
  5.09663889]
The cluster 2 centre are found to be [18.71966667 16.2995      0.884745    6.20988333  3.72128333  3.61626667
  6.06386667]
The cluster 3 centre are found to be [14.65731343 14.47283582  0.87820299  5.57362687  3.27565672  2.66252537
  5.19283582]
For cluster 1, maximum datapoint distance is 3.8143674445694598
For cluster 1, minimum datapoint distance is 0.2515151088287457
For cluster 1, variance is 0.515203164441043
For cluster 2, maximum datapoint distance is 3.4381895247228407
For cluster 2, minimum datapoint distance is 0.216254495707924
For cluster 2, variance is 0.5851616732317656
For cluster 3, maximum datapoint distance is 3.1840272408261536
For cluster 3, minimum datapoint distance is 0.17121919536174962
For cluster 3, variance is 0.5513951592094533
The labels for each of them are [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0 2 2 0 2 2 2 2 2 2 0 2 2 2 2 2 2 2 2 2 1 2
 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0 0 0 2 2 2 2 0 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 2 2 2 1 2 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
The Silhouette score is 0.4727232858211616
Considering points outside 90% radius of cluster as outliers:
For cluster 1, outliers are 1 out of 72
For cluster 2, outliers are 3 out of 60
For cluster 3, outliers are 3 out of 67
```

c). Randomly created data: DBSCAN algorithm can be tested on a random set of points that were generated as shown in Fig 3. This data has some noise added to it. These act as the outliers in this data.

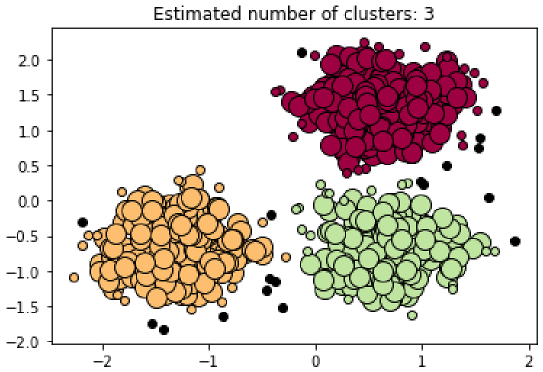
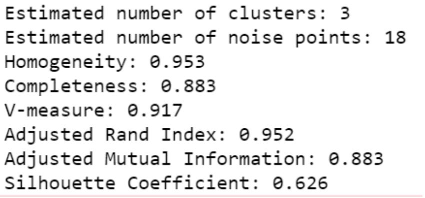


Fig 3. Clusters for randomly created dataset

The generated summary:



```
Estimated number of clusters: 3
Estimated number of noise points: 18
Homogeneity: 0.953
Completeness: 0.883
V-measure: 0.917
Adjusted Rand Index: 0.952
Adjusted Mutual Information: 0.883
Silhouette Coefficient: 0.626
```

Conclusion: In this paper, we proposed a summarization method that associates description with the clusters obtained by different models, to help us evaluate or interpret the clusters. As shown in the experimentation part, the proposed method finds intuitive, eloquent labels that describe well or "verbalize" these clusters. This method is fully applicable to various datasets, including continuous attributes, missing values, outliers, and can be a new, in-depth, and consistent tool for cluster interpretation/evaluation. Using this description, one can relate certain things associated with the data results. For e.g., the reason for data points to lie within a particular cluster or which feature distinguishes a data point to lie in a given cluster. Such type of verbalization helps in understanding the underlying features in the dataset.

References

[1] Chris Ding and Xiaofeng He. 2004. K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery, New York, NY, USA, 29. DOI:https://doi.org/10.1145/1015330.1015408.

[2] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 226–231.C. Qi, L. Xiao, T. Wang, J. Li, and L. Li, "A highly reliable memory cell design combined with layout-level approach to tolerant single-event upsets," *IEEE Trans. Device Mater. Rel.*, vol. 16, no. 3, pp. 388–395, Sep. 2016.

[3] Al- Zoubi, Moh'd Belal & Raw, Mohd. (2008). An Efficient Approach for Computing Silhouette Coefficients. Journal of Computer Science. 4. 10.3844/jcssp.2008.252.255..

[4] Y. Kameya, S. Nakamura, T. Iwasaki and T. Sato, "Verbal Characterization of Probabilistic Clusters Using Minimal Discriminative Propositions," 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, Boca Raton, FL, 2011, pp. 873-875, doi: 10.1109/ICTAI.2011.136.