Analysis of Clustering Algorithms Through Image Recognition

Ben Barber, Jordan Boyle, Axel Arellano
Option: B

September 2020

Technical

1. **What is the problem or technique that you plan to analyze?** The machine learning technique chosen for analysis is Clustering. We will be comparing three different algorithms: Hierarchical Agglomerate Single Link, K-means, and DBSCAN. The metrics of these algorithms will be compared through their performance with image recognition. 10,000 labeled images from a camera trap will be used to identify different types of animals in Eastern North America. From that dataset there are 23 different types of animals labeled. The results from these identifications will be used to compare the the three algorithms.

2. **What properties would you like to analyze/prove about this problem or technique?**

   The analysis will evaluate and compare accuracy, scalability, and the runtime and memory efficiency of Hierarchical Agglomerative Single Link Clustering, K-Means, and DBSCAN. The four proceeding metrics will be evaluated for differing datasets. These datasets will differ from 4-23 different dimensions. To account for each algorithms scalability, we will be using 4 different sized datasets. The datasets will contain 2500, 5000, 7500, 10000 images each. By comparing the results from these tests, we will be able to find which algorithm performs best given certain conditions.

3. **Cite at least 6 related papers that you plan to review.**

   I. Comparative Study of Clustering Based Colour Image Segmentation Techniques [1].

   II. An Overview of Clustering Models with an Application to Document Clustering [2].

   III. An empirical comparison of Clustering using hierarchical methods and K-means Clustering [3].

   IV. Anomaly Detection via Correlation Clustering [4].

   V. Clustering Techniques and Research Challenges in Machine Learning [5].

   VI. Predicting the Accuracy of Fractional of Patron's Activities in Online Social Networks Using Novel K-Means Clustering Algorithm Comparing with Agglomerative Hierarchical Clustering Algorithm [6].

Management

The intended dataset is the ENA24-detection Dataset. [7].
In order to accomplish the goal of analyzing these three algorithms and the metrics previously mentioned, the following steps will be carried out:

1. **Researching Clustering Techniques and Measures:**
   First we begin by thoroughly researching the three cluster techniques: Hierarchical clustering, K-Means, and DBSCAN. Understand the principles behind each algorithm, including how they partition data and the parameters they require. The weighting scheme needed for the distances between clusters will be identified so the data can be further annotated.

2. **Split Dataset for Scalability/Dimension Testing:**
   The dataset will be split into four differing sizes so scalability can be later tested. The data will also be split into different groups of varying sizes, and further divided into training and testing folds.

3. **Training the Clustering Models**
   The three models (Hierarchical clustering, K-Means, DBSCAN) will be trained on the split datasets. Using the appropriate parameters and settings for each algorithm, including the number of clusters and any distance measures specified.

4. **Testing the Clustering Models**
   The trained clustering models will be evaluated using the testing datasets. Apply each model to the test data to obtain cluster assignments or labels for the data points. Metrics for each test will be recorded.

5. **Comparison of Results**
   Compare the performance of the clustering models based on accuracy, scalability, and the efficiency of the algorithms. Identify the strengths and weaknesses of each model in the different scenarios.

References

[1] Samira Chebbout and Hayet Farida Merouani. Comparative study of clustering based colour image segmentation techniques. In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*, pages 839–844, 2012.

[2] Iva Pauletic, Lucia Nacinovic Prskalo, and Marija Brkic Bakaric. An overview of clustering models with an application to document clustering. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2019.

[3] P. Praveen and B. Rama. An empirical comparison of clustering using hierarchical methods and k-means. In *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, August 2016.

[4] Peter Shaw, Joseph R. Barr, and Faisal N. Abu-Khzam. Anomaly detection via correlation clustering. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 307–313, 2022.

[5] Ritesh C. Sonawane and Hitendra D. Patil. Clustering techniques and research challenages in machine learning. In *2020 Fourth International Conference on Computing Methodologies and Communication(ICCMC)*, March 2020.

[6] V. Venkatesh and A. Shri Vindhya. Predicting the accuracy of fractionation of patron's activities in online social networks using novel k-means clustering algorithm comparing with agglomerative hierarchical clustering algorithm. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Septemeber 2023.

[7] Hayder Yousif, Roland Kays, and Zhihai He. Dynamic programming selection of object proposals for sequence-level animal species classification in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.