Geometrical Homogeneous Clustering for Image Data Reduction

Shril Mody * 1  Janvi Thakkar * 1  Devvrat Joshi * 1  Siddharth Soni * 1  Rohan Patil 1  Nipun Batra 1

Abstract

In this paper, we present novel variations of an earlier approach called homogeneous clustering algorithm for reducing dataset size. The intuition behind the approaches proposed in this paper is to partition the dataset into homogeneous clusters and select some images which contribute significantly to the accuracy. Selected images are the proper subset of the training data and thus are human-readable. We propose four variations upon the baseline algorithm-RHC. The intuition behind the first approach, RHCKON, is that the boundary points contribute significantly towards the representation of clusters. It involves selecting \( k \)-farthest and one nearest neighbour of the centroid of the clusters. In the following two approaches (KONCW and CWKC), we introduce the concept of cluster weights. They are based on the fact that larger clusters contribute more than smaller sized clusters. The final variation is GHCIDR which selects points based on the geometrical aspect of data distribution. We performed the experiments on two deep learning models- Fully Connected Networks (FCN) and VGG1. We experimented with the four variants on three datasets- MNIST, CIFAR10, and Fashion-MNIST. We found that GHCIDR gave the best accuracy of 99.35%, 81.10%, and 91.66% and a training data reduction of 87.27%, 32.34%, and 76.80% on MNIST, CIFAR10, and Fashion-MNIST respectively.

1. Introduction

Various approaches for dataset reduction have been proposed in the past. One such popular technique is called homogeneous clustering (RHC) (Ougiaroglou & Evangelidis, 2012). RHC clusters the given dataset by considering the labels of each data point. The dataset is partitioned into homogeneous clusters; i.e. all the data points belonging to that cluster will have the same label. The final reduced dataset is constructed by taking the centroids of these homogeneous clusters.

The RHC algorithm only includes the centroid of the homogeneous clusters in the reduced set. It disregards the importance of other images which forms the cluster boundary. Boundary points can potentially play a significant role in distinguishing the labels of neighbouring clusters. The varying size of the clusters is also not taken into account by RHC. Larger clusters contain more images, and so their contribution should be more. Our work attempts to further improve the idea of RHC by taking into account the geometric features involved in generating clusters. We focused on reducing the dataset such that it stays human readable. Human-readable data can be used to interpret the results in case of unexpected output easily.

We propose four variations of our approach. All the algorithms differ in the selection of data points after the creation of homogeneous clusters. The first approach, RHCKON focuses on picking the data points greedily near the centroid of the clusters (\( k \)-farthest and 1-nearest neighbours). The intuition behind RHCKON is that the boundary points are a factor for classification between adjacent clusters (Olvera-López et al., 2010). The second variation, KONCW gives weights to each cluster according to their size and selects images based on these weights. Large-sized clusters could contribute more towards the accuracy as they contain more images, so we can include them by giving weights. The next variation, CWKC greedily chooses the images which are evenly spread across the cluster boundary. Here, the number of images chosen from the cluster depends on their weight. The final algorithm, GHCIDR combines all the previous variations by dividing the cluster into annular regions and selecting images from each region. We can get a notion of the entire cluster by selecting images equally spaced from the centroid rather than selecting only the farthest and nearest images from a cluster.

All these four variants were tested on three image datasets: MNIST (60,000 training images) (Deng, 2012), CIFAR10 (50,000 training images) (Krizhevsky et al., 2009) and Fashion-MNIST (60,000 training images) (Xiao et al., 2017).
We tested our reduced datasets on two deep learning models, namely, Fully Connected Networks (FCN (Long et al., 2015)) and VGG1 (Wang et al., 2015). We found that all four variants performed better than the baseline RHC algorithm on both models. GHCIDR on VGG1 performs the best among the four and gives approximately the same accuracy as that of the model trained on full dataset. +

Our code ¹ is fully reproducible.

2. Background (Reduction through Homogeneous Clustering)

A variety of sampling-based algorithms were proposed to reduce the size of the annotated datasets. In this section, we describe one of the previously proposed approaches called “Reduction through Homogeneous Clustering (RHC)”. RHC focused on constructing homogeneous clusters. Homogeneous clusters contain data points belonging to the same class. The algorithm maintains a set of clusters. Initially, this set consists of the proper training dataset as a single cluster. On each iteration, we pop out a cluster from this set. Now, there are two possible conditions for the clusters: homogeneous & non-homogeneous. If the cluster is homogeneous, it gets added to the Condensed Set. Condensed Set is the final output of the RHC algorithm. If it is non-homogeneous, then we recursively cluster the points using $K$-means (Chen & Xia, 2009) algorithm until either all clusters become homogeneous or there is no cluster left.

At each step, we initialize $K$-means by the centroids of each class belonging to the cluster. The dataset is reduced by selecting centroids of individual homogeneous clusters. RHC was tested on various annotated datasets (Alcalá-Fdez et al., 2011), namely, Letter recognition (LR), Pen-Digits (PD) and Landsat Satellite (LS). It gave better accuracy with lesser time complexity than classical ML algorithms such as K-NN and Prototype Selection by Clustering (PSC) (Olvera-López et al., 2010).

The original RHC work primarily focused on reducing the annotated dataset. One of the shortcomings of the RHC algorithm was that it selected centroids of the images that could be non-human readable as shown in Figure 1. We modified the RHC algorithm to produce human-readable samples giving high accuracy.

3. Proposed Approach

The RHC algorithm chooses the centroid of each homogeneous cluster. This aggregation of images is not human readable. RHC also ignores the size of clusters and selects a single image (centroid), making the contribution of small and large clusters to be equal in the reduced dataset. These drawbacks are caused by RHC’s selection method of images from homogeneous clusters. We propose four different variants which overcome these limitations.

Figure 1. Non Human readable images of RHC (left: CIFAR10 (class: bird), right: MNIST (class: 9))

3.1. Approach 1: RHC + $K$-Farthest + OneNearest (RHCKON)

The RHC algorithm does not select the boundary images from each of the homogeneous clusters, which can potentially be the distinguishing factor for classification. Boundary points of a cluster Most of the boundary points have properties similar to more than one cluster, so classifying them correctly can have a considerable impact on test accuracy.

Algorithm: In RHCKON (Figure 2), we choose one nearest point and $K$ farthest points from the centroid of each cluster. Here, these $K$ points represent the boundary points for each cluster.

In RHCKON, we envision the following three main drawbacks.

1. It may be prone to outliers (more boundary points in the reduced set).
2. Reduction is less in comparison to RHC because of more images selected from each cluster.
3. RHCKON selects the same number of points from each cluster. Therefore, it could lead to more fraction of images from smaller sized clusters in the reduced set.

To illustrate the point 3 above, we plot the size of clusters versus cluster count on CIFAR10 dataset in Figure 3. The clusters of smaller size are significantly large in number, whereas clusters of larger size are small in number. Out of 12,400 clusters, nearly 10,000 clusters lie in a size range between 1 and 5. However, the significant chunk of data points lies in the large-sized clusters. Thus, by selecting one nearest and $K$ farthest samples, we are giving the same weightage to all the clusters irrespective of their size.

¹Here is the link to the code https://github.com/SoniSiddharth/GHCIDR.
Figure 2. Approach 1: RHCKON. This image shows one of the many homogeneous clusters. We sample the point closest to the centroid and K farthest points from the centroid. Choosing boundary points may provide diverse images to achieve better generalization.

Figure 3. No. of Clusters vs. Cluster Size. Clusters with small size are large in number, and their count decreases exponentially with their size. 10000 smaller clusters contain 15000 images whereas 2400 larger clusters contain the remaining 35000 images.

3.1.1. Approach 2 - RHCKON+Cluster Weightage (KONCW)

In our second approach (shown in Figure 4) instead of selecting an equal number of images via RHCKON, we can give a weightage (fraction of selected images) to each cluster according to their size. Thus, the largest cluster will have the highest weightage and vice versa.

Algorithm: We need to select at least one image from each cluster to have a representation of these isolated clusters in our reduced dataset.

- Select a fraction of reduction ($\alpha$), i.e. select from each cluster a fraction of points $(1 - \alpha)$.
- $\max((1 - \alpha) \times \text{size of cluster}, 1)$ is the number of images selected from each cluster.
- Let the number of the images from the $i^{th}$ cluster be $X_i$.

- Select one nearest image and, $X_i - 1$ farthest images from the the $i^{th}$ cluster.

However, we envision the following limitations.

1. Larger clusters have higher weightage in the reduced set. Due to this, the fraction of boundary points selected from these clusters increases with the size of clusters which increases the probability of outliers in our reduced dataset.

2. Assume an N-dimensional ball representing the cluster of data points. We are choosing the farthest points irrespective of the direction from the centroid of the ball/cluster. Thus, there is a considerable probability that the selected farthest points might be in the same direction from the centroid of the cluster.

3.1.2. Approach 3 - KONCW+K-Centre (CWKC)

We need to select points in different directions from the cluster’s centroid. Figure 5 illustrates the CWKC approach.

Algorithm:

- Initialize the set of centres with one nearest and one farthest point.
- Choose the next point, which is farthest from all the points in the set of centres.
- It is equivalent to finding a point in the cluster whose minimum distance from all the points in the set is maximum.
- Continue extending the set until its size is less than the allowed points from the cluster (weightage).
- Apply the same algorithm on all the clusters, and we get our condensed dataset.
3.1.3. Approach 4 - Geometrical Homogeneous Clustering for Image Data Reduction (GHCIDR)

Selecting the nearest point from the centroid makes a notion of centrality that the given point represents the entire cluster. In our final approach called GHCIDR (shown in Figure 6) we choose images uniformly from the entire volume of the cluster. This way, we can encompass all the cluster features.

Algorithm: Consider a cluster as a ball of $N$ dimensions.

- Let $maxDist$ be the distance of the farthest point from the centroid.
- Divide $maxDist$ into partitions of equal size to get $(maxDist/\beta)$ annuluses of the ball, with the innermost part being a sphere.
- $\beta$ depends on the weightage of the cluster and the reduction rate $\alpha$.
- For each annulus $i$, let the inner radius and outer radius of the annulus be $R_{1i}$ and $R_{2i}$.
- Find the point belonging to annulus $i$ and having distance from the centroid to be $(R_{1i} + R_{2i})/2$.
- Put that point along with the one nearest point from the central sphere into the condensed set.

We set $\beta$ to get exactly one point from each annulus.

$$\beta = \frac{maxLength}{((1-\alpha) \times (size\ of\ cluster))}$$

4. Evaluation

4.1. Datasets

MNIST Dataset: It is collection of 60,000 handwritten images with 10 classes. Dimensions of each image are 28x28.

FMNIST Dataset: It is collection of 60,000 clothes images with 10 classes. Dimension of each image are 28x28.

CIFAR10 Dataset: It is a dataset of 50,000 coloured images with 10 classes. Dimension of each image are 32x32x3. There is a larger version of this dataset with 100 classes called CIFAR100.

4.2. Metrics

Accuracy: We are using accuracy as a metric to compare the experimental results.

4.3. Experimental Settings

- We experimented on two deep learning algorithms are FCN (Fully Connected Network), and VGG1. We used Nvidia P100 GPU with 12 GB RAM for dataset reduction and training.
- Four variants and the original RHC algorithm were applied on the training datasets to obtain the reduced dataset.
- For FCN and VGG1, we used a batch size of 32 and 64, and epochs equal to 5 and 100, respectively.
- We sampled a random subset of images whose size was equal to the reduced dataset. For example in Table 1, RHCKON on MNIST reduces dataset size by 90.13% or samples 9.87% of the dataset. We select 9.87% of the points randomly in the random sampling baseline. We report the average accuracy on randomly sampled datasets was calculated by varying the random seed.
- We evaluated the testing datasets and reported the accuracy for models trained on reduced data and random data.
Geometrical Homogeneous Clustering

| Algorithm   | MNIST %Accuracy on reduced data | %Reduction | MNIST %Accuracy on random data | %Reduction | FMNIST %Accuracy on reduced data | %Reduction | FMNIST %Accuracy on random data |
|-------------|---------------------------------|------------|--------------------------------|------------|---------------------------------|------------|---------------------------------|
| Full Dataset| 97.35                           | 0.00       | -                              | 87.41      | 0.00                            | -          | -                              |
| RHC         | 90.03                           | 95.06      | 93.14                          | 64.34      | 90.93                           | 82.52      |
| RHCKON      | 96.02                           | 90.13      | 94.36                          | 75.91      | 81.86                           | 84.24      |
| KONCW       | 96.70                           | 74.05      | 96.10                          | 82.40      | 69.07                           | 84.10      |
| CWKC        | 95.72                           | 79.00      | 95.70                          | 82.06      | 78.14                           | 83.92      |
| GHCIDR      | 96.83                           | 87.27      | 94.41                          | 83.96      | 76.80                           | 82.88      |

Table 1. Accuracy on reduced data, random data and % reduction of MNIST and FMNIST datasets trained on FCN model for all four variants and RHC.

| Algorithm   | MNIST %Accuracy on reduced data | %Reduction | MNIST %Accuracy on random data | %Reduction | FMNIST %Accuracy on reduced data | %Reduction | FMNIST %Accuracy on random data |
|-------------|---------------------------------|------------|--------------------------------|------------|---------------------------------|------------|---------------------------------|
| Full Dataset| 99.51                           | 0.00       | -                              | 93.25      | 0.00                            | -          | -                              |
| RHC         | 97.35                           | 95.06      | 98.00                          | 75.39      | 90.93                           | 87.86      |
| RHCKON      | 99.14                           | 90.13      | 98.70                          | 88.60      | 81.86                           | 89.51      |
| KONCW       | 99.46                           | 74.05      | 99.12                          | 90.96      | 69.07                           | 90.18      |
| CWKC        | 99.21                           | 79.00      | 98.98                          | 89.74      | 78.14                           | 90.13      |
| GHCIDR      | 99.35                           | 87.27      | 98.16                          | 91.66      | 76.80                           | 90.28      |

Table 2. Accuracy on reduced data, random data and % reduction of MNIST and FMNIST datasets trained on VGG1 model for all four variants and RHC.

4.4. Results

From Table 1, we can observe that:

- Accuracy of full dataset of MNIST when trained on FCN is 97.35%. When reduced with GHCIDR, it gave an accuracy of 96.83% and a reduction of 87.27%.
- Accuracy of full dataset of Fashion-MNIST, on FCN is 87.41%. When reduced with GHCIDR, it gave an accuracy of 83.96% and a reduction of 76.80%.

From Table 2, we can observe that:

- Accuracy of full dataset of MNIST on VGG1 is 99.51%. When reduced with GHCIDR, it gave an accuracy of 99.35% and a reduction of 87.27%.
- Accuracy of full dataset of Fashion-MNIST on VGG1 is 93.25%. When reduced with GHCIDR, it gave an accuracy of 91.66% and a reduction of 76.80%.

From Table 3, we can observe that:

- Accuracy of full dataset of CIFAR10 on VGG1 is 82.87%. When reduced with GHCIDR, it gave an accuracy of 81.10% and a reduction of 32.34%.

4.5. Analysis

All four approaches reduced the dataset efficiently and gave an accuracy similar to that of the full dataset. They all performed better than the baseline RHC in terms of accuracy. GHCIDR outperformed all the variants for every dataset in all experiments. For the FCN model on MNIST and FMNIST datasets, KONCW gave a higher accuracy than RHCKON and CWKC but with a reduced % reduction. Because cluster weights in KONCW, ensuring that the selection of images was proportional to the size of clusters. For the same amount of reduction, random data also gave less accuracy than KONCW. GHCIDR gave more reduction than KONCW and more accuracy because it samples images evenly from the cluster. KONCW inculcates the advantages of boundary points which help to increase the classification but decreases the reduction. On the other hand, GHCIDR takes advantage of KONCW and CWKC and considers the geometrical aspect of the data distribution. This leads to increased overall accuracy while keeping high % reduction for the GHCIDR algorithm. By tuning the GHCIDR parameters - $\alpha$ & $\beta$, we can achieve more accuracy but at the cost of a lesser reduction rate. This leads to a tradeoff between accuracy and reduction rate.
Geometrical Homogeneous Clustering

| Algorithm    | %Accuracy on reduced data | %Reduction | %Accuracy on random data |
|--------------|----------------------------|------------|--------------------------|
| Full Dataset | 82.87                      | 0.00       | -                        |
| RHC          | 67.59                      | 75.67      | 70.8                     |
| RHCKON       | 76.04                      | 51.34      | 75.24                    |
| KONCW        | 77.80                      | 47.33      | 77.93                    |
| CWKC         | 76.24                      | 42.37      | 78.26                    |
| GHCIDR       | **81.10**                  | **32.34**  | **79.57**                |

Table 3. Accuracy on reduced data, random data and % reduction of CIFAR10 datasets trained on VGG1 model for all four variants and RHC.

5. Conclusion and Future Work

We proposed novel variations to the RHC algorithm in this work. The best performing approach called GHCIDR was based on the intuition that we can get representative and diverse samples by sampling “throughout” the volume of the cluster. GHCIDR performed better than the proposed approaches, random baseline and the original RHC.

In the future, we plan to compare this work with non-RHC algorithms like Light-weight coresets (Bachem et al., 2018), and other sampling algorithms. And we also plan to try different distance metrics such as SSIM (Bae & Kim, 2015) and image augmentation techniques to modify the current algorithm. We also plan to reduce Imagenet (Deng et al., 2009) and similar large datasets using the proposed approaches.

References

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2011.

Bachem, O., Lucic, M., and Krause, A. Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1119–1127, 2018.

Bae, S.-H. and Kim, M. A novel ssim index for image quality assessment using a new luminance adaptation effect model in pixel intensity domain. In *2015 Visual Communications and Image Processing (VCIP)*, pp. 1–4. IEEE, 2015.

Chen, Z. and Xia, S. K-means clustering algorithm with improved initial center. In *2009 Second International Workshop on Knowledge Discovery and Data Mining*, pp. 790–792. IEEE, 2009.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Olvera-López, J. A., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. A new fast prototype selection method based on clustering. *Pattern Analysis and Applications*, 13(2):131–141, 2010.

Ougiaroglou, S. and Evangelidis, G. Efficient dataset size reduction by finding homogeneous clusters. In *Proceedings of the Fifth Balkan Conference in Informatics*, pp. 168–173, 2012.

Wang, L., Guo, S., Huang, W., and Qiao, Y. Places205-vggnets models for scene recognition. *arXiv preprint arXiv:1508.01667*, 2015.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.