A Quartile-based Initialization for Deep Embedded Clustering on News Articles

Jafor Mohammad
Department of Computer Science
BRAC University
Email: jafor.mohammad@g.bracu.ac.bd

***Abstract*—In this paper, we discuss a technique for clustering centroid initialization for Deep Embedded Clustering (DEC), based on the statistics of the quartile. We apply our approach to the AG News dataset, taking Sentence-BERT embedding to show that our initialization technique is an alternative to the common methods, such as KMeans++, for initialization of centroids. Although our approach can support clustering performance almost as well as DEC with KMeans++ initialization (0.9858 Silhouette Score and 0.0189 Davies-Bouldin Index), we note that the generalization and possible benefits of our method require additional testing on other datasets and other such algorithms.**

***Index Terms*—Deep Embedded Clustering, Quartile-based Initialization, KMeans++, AG News Dataset, Sentence-BERT, Clustering.**

I. Introduction

Clustering in high-dimensional text data is still a hard task, particularly if such embeddings are trained in an unsupervised way. Deep Embedded Clustering (DEC) combines the principles of autoencoding and clustering, in order to learn a compact latent representation that is appropriate for the problem of clustering. Initialization of cluster centroids is an important aspect of DEC's success. We present a new initialization technique in this study, based on statistical quartiles of the latent space that would aid in increasing the robustness and performance of DEC. We test our method on the AG News dataset using Sentence-BERT and demonstrate that it performs comparably with traditional k-means initialization. Nevertheless, we recognize that further testing with a variety of datasets and clustering methods is required to completely assess its viability.

II. Related Work

Many clustering algorithms, like the K-Means [2], are used for clustering text files as a rule, but are not quite suitable for work in high-dimensional spaces, especially if embeddings are performed without supervision. Although K-Means++ overcomes random initialization and probabilistically selects initial centroids [3], it continues to struggle with complicated latent spaces. DEC [1] suggests a joint learning model for feature representation and cluster assignment with the initialization of K-means in latent space. Our efforts are an extension of DEC, bringing in a deterministic and statistically rooted initialization scheme via quartiles, providing a new take on centroid initialization. Although our technique is not guaranteed to perform better than K-Means++, our technique is an alternative that deserves more exploration in terms of data and clustering algorithms.

III. Dataset and Preprocessing

We take the AG News dataset that includes news articles labeled as about the four classes: World, Sports, Business, and Sci/Tech. The dataset was accessed from the HuggingFace fancyzhx/ag_news repository. We filtered out short texts (length ¡ 50 characters) to make sure we have a meaningful semantical content for further embedding. Paraphrase-MiniLM-L12-v2 384-dimensional embeddings generated by Sentence-BERT [4] were used, which is appropriate in terms of performance as well as computation efforts. Those embeddings were standardized with the help of StandardScaler [5] to normalize the distribution of features and ensure stable convergence of models during training.

IV. Methodology

A. Autoencoder Architecture

We developed a completely connected symmetric autoencoder with an input of 384, having a latent dimension of 32. There are three layers that make up the encoder: $384 \Rightarrow 512 \Rightarrow 256 \Rightarrow 32$, with ReLU activation. The decoder mirrors the encoder: $32 \rightarrow 256 \rightarrow 512 \rightarrow 384$. This configuration provides a gradual reduction in dimensionality, keeping the information the same, only reducing the size of the embedding space. There was empirical experimentation used to arrive at 32 as the latent dimension that aimed at striking a balance between compression and fidelity to clustering. We pretrain the autoencoder for 10 epochs based on the Adam optimizer with learning rate = 0.001 and MSE loss. A total of approximately 673,696 trainable parameters are present in the model.

B. Hyperparameter Optimization

Empirical testing on the training data was used to choose the hyperparameters, such as latent size, learning rate, and epochs. The latent dimension of 32 was a good tradeoff of representational power and clustering separability. We noticed that lower than 16 damaged clustering quality, and going beyond 64 introduced additional complications. Learning rates were tuned based on the observation of the loss convergence, where 0.001 for pretraining and 0.0005 for the clustering phase produced stable results. The batch size was tuned to 256 to make the calculation efficient.

C. Regularization and Normalization

There is no batch normalization or dropout layers used explicitly because the initial experiments have shown that they do not provide any significant improvement in clustering performance. Furthermore, based on the rather shallow network and low risk of overfitting in the case of unsupervised training, L2 regularization was not used. Embeddings, however, were normalized using z-score normalization to obtain consistency in the scaling of features.

D. Quartile-Based K-Means Initialization

Once pretrained, we push all of the data through the encoder to get latent representations. We use the first quartile (Q1), median (Q2), the third quartile (Q3), and the mean of the latent dimensions to construct four representative cluster centroids, instead of random initialization or KMeans++. This deterministic approach models the statistical structure of the data and avoids it from suffering from randomness. These values are then given as initializations of k-means [2] with `n_init=1` and `max_iter=300`. The obtained centroids are allocated to the DEC clustering layer.

E. Deep Embedded Clustering (DEC)

The autoencoder is fine-tuned alongside the clustering layer by DEC. We use the KL divergence loss between the soft cluster assignment q and a target distribution p made from q. Training of the model is carried out on 30 epochs using Adam with a reduced learning rate of 0.0005. The clustering layer makes use of the centroids that have been initialized by our quartile-based approach. The KL loss encourages positive assignment and compact clusters.

V. Results

The Silhouette Score value of the AG News dataset was 0.9858, and the Davies-Bouldin Index value was 0.0189. Such metrics imply that the clustering is highly cohesive and well separated. But when compared to DEC initialized with KMeans++, our method with the quartile initialization gives similar clustering performance. A t-SNE demonstrates further that both approaches have clustering that is well separated.

A. t-SNE Visualizations

We present t-SNE plots to visually compare the effectiveness of different initialization strategies:

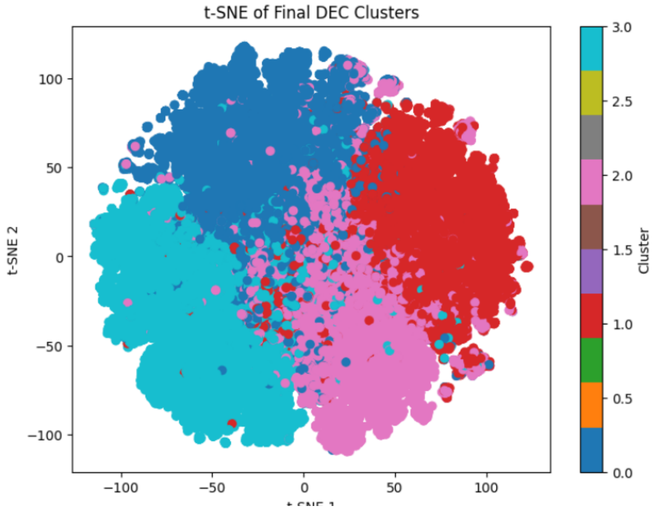

Fig. 1. t-SNE plot for clusters initialized with random initialization. (Silhouette Score: 0.0593, Davies-Bouldin Index: 3.6059)

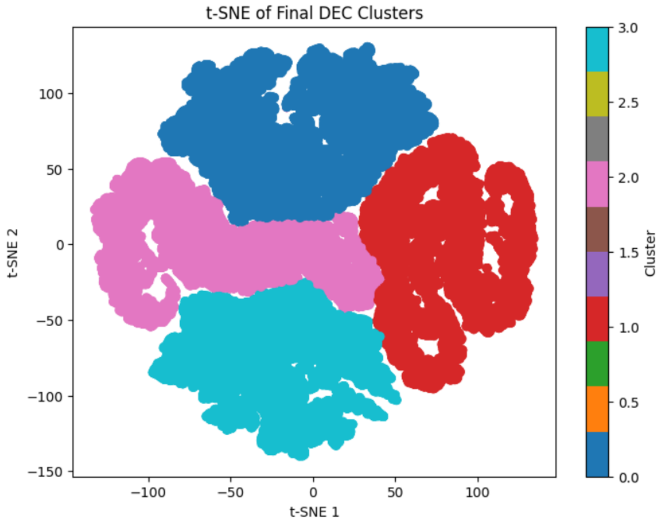

Fig. 2. t-SNE plot for clusters initialized with KMeans++. (Final Silhouette Score: 0.9819, Davies-Bouldin Index: 0.0257)

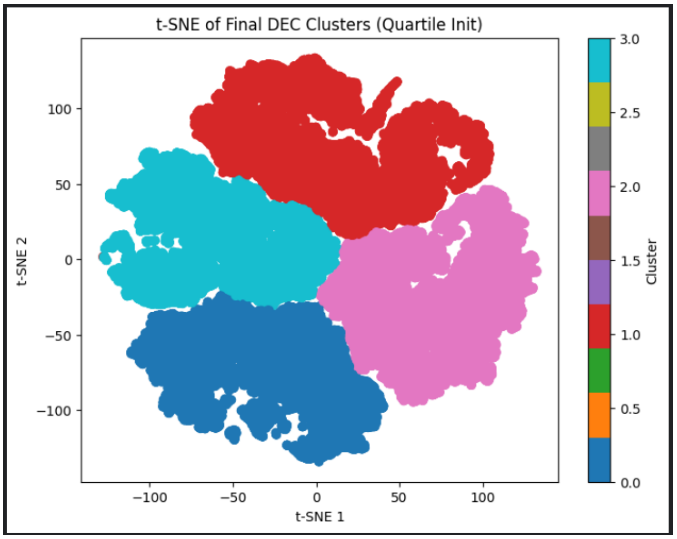

Fig. 3. t-SNE plot for clusters initialized with the quartile-based method (proposed). (Final Silhouette Score: 0.9858, Davies-Bouldin Index: 0.0189)

B. Comparative Analysis

From the t-SNE visualizations, we can see that there are similar clustering behaviours between our quartile-based initialization and KMeans++. Whilst the clustering can be differentiated in both cases, the difference is negligible indicating that there is little difference between the methods in this dataset. Our approach's strength stems from its deterministic nature and might provide better replications across different runs as compared to the K-Means++ method.

VI. Limitations

Although our quartile-based initialization approach provides similarly good clustering results, it relies on, in the sense of the global structure of the latent space – adequate quartile statistics. This assumption may not apply to datasets with distributions that are highly skewed or multimodal. Further on, a significant drawback of the quartile-based initialization method is also its scalability in terms of the number of features in the dataset. The time consumed in the computation of quartiles for individual features would increase as the number of columns (features) in the datasets increases proportionally. This leads to longer preprocessing and initialization times, which makes the method less efficient for very high-dimensional data.

VII. Future Work

Future studies should be aimed at the evaluation of the method of initialization employed in quartiles over a wider range of datasets, especially those that differ in size and dimension of features, and complexity of clusters. This will also enable us to better understand its robustness and adaptability. Additionally, additional research will limit itself to clustering algorithms that require initial cluster centers for their execution, thus omitting algorithms that do not need starting points. This focus provides relevance and applicability of the initialization strategy under consideration.

VIII. Conclusion

We proposed a quartile-based initialization method for DEC; providing a potential alternative to old techniques such as K-Means++. The results we obtained on the AG News dataset demonstrate comparability of the results, meaning that the method is suitable for the clustering of textual data. However, we do stress that in order to fully assess the strengths and potential of this approach, it is required to conduct further experiments on other sets of data and clustering algorithms.

References

[1] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[3] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2007, pp. 1027–1035.

[4] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[5] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.