E0:270 - Machine Learning - Clustering using Deep Learning

Mayank Singh [1] **Shivank Gupta** [1] **Tapan Bhardwaj** [1]

Abstract

Clustering is central to many data-driven application domains and has been studied extensively in terms of distance functions and grouping algorithms. Relatively little work has focused on learning representations for clustering. In this project, we have explored recent deep learning approaches like Improved Deep Embedding Clustering(Xifeng Guo, 2017), Towards K-means-friendly Spaces(Yang et al., 2016) and Variational Deep Embedding(Jiang et al., 2016). Our objective is to implement these three approaches and test their performance on one synthetic and two real world data.

1. Problem Statement

Unsupervised clustering is a vital research topic in data science and machine learning. However, when the dimension of input feature space (data space) is very high, the clustering becomes ineffective due to unreliable similarity metrics. Transforming data from high dimensional feature space to lower dimensional space in which to perform clustering is an intuitive solution and has been widely studied.

2. Literature review

We have gone through the following Deep Learning architecture for Clustering.

2.1. DEC : Deep Embedded Clustering(Xie et al., 2016)

This method simultaneously learns feature representations and cluster assignments using deep neural networks. It learns a mapping from the data space to a lower-dimensional feature space in which it iteratively optimizes a clustering objective without including the reconstruction cost.

[1]Department of Computer Science and Automation, Indian Institute of Science, Bangalore. Correspondence to: Mayank Singh <mayanksingh@iisc.ac.in>, Shivank Gupta <shivankgupta@iisc.ac.in>, Tapan Bhardwaj <tapanb@iisc.ac.in>.

Midterm Project Report, E0:270, Machine Learning, 2018.

2.2. Towards K-means-friendly Spaces(Yang et al., 2016)

The paper introduces us to a Deep Clustering Network which models the relationship between observable data and its k-mean clustering friendly latent representation. It contains of two halves, the first half converts the observable data of high dimensionality into clustering friendly latent representation of low dimension, and the second half reconstructs the latent vector (output of first half) close to the original data of high dimensionality. This helps in avoiding trivial solutions. The cost function consists of reconstruction as well as clustering losses. Initialization of network parameters is done using layer-wise pre-training(Bengio et al., 2007) Optimization is done in subproblems keeping either Network Parameters or Clustering Parameters as constant.

2.3. IDEC : Improved Deep Embedding Clustering(Xifeng Guo, 2017)

Some pioneering work proposes to simultaneously learn embedded features and perform clustering by explicitly defining a clustering oriented loss.But considering only clustering loss may corrupt feature space and which can lead to poor clustering performance. To address this issue, Improved Deep Embedded Clustering (IDEC) algorithm take care of data structure preservation. To maintain the local structure of data generating distribution, an auto-encoder is applied. By integrating the clustering loss and autoencoders reconstruction loss, IDEC can jointly optimize cluster labels assignment and learn features that are suitable for clustering with local structure preservation.

2.4. VaDE : Variational Deep Embedding(Jiang et al., 2016)

In this paper a clustering framework combining Variational Auto-encoder(Kingma & Welling, 2013) and Gaussian Mixture Model is given. The VaDE has two subnetworks:

(a) Encoder network which converts observable data X into latent embedding Z based on Gaussian Mixture Model.

(b) Decoder Network which takes latent representations Z

and converts it to observable data X.

This is unsupervised generative approach to clustering where one can assume K clusters for implementing Gaussian Mixture Model. The encoder network is used to maximize evidence lower bound of VaDE which contains reconstruction term and KL-divergence term of mixture of Gaussian.

3. Dataset Description

MNIST: The MNIST dataset consists of 70000 handwritten digits. The images are centered and of size 28 by 28 pixels. We reshaped each image to a 784- dimensional vector.

HHAR: The Heterogeneity Human Activity Recognition (HHAR) dataset contains 10299 sensor records from smart phones and smart watches. All samples are partitioned into 6 categories of human activities and each sample is of 561 dimensions.

Synthetic Data: We will generate four clusters, each of which has 2,500 samples in 2 dimension. This two-dimensional domain is a latent domain which we do not observe, What we observe is $x_i \in R^{100}$ that is obtained via the following transformation: where $W \in R^{10\times 2}$ and $U \in R^{100\times 10}$ are matrices whose entries follow the zero-mean unit-variance i.i.d. Gaussian distribution and $h_i \in R^2$ which is our original data point in 2 dimension.

$$x_i = \sigma(U\sigma(Wh_i))$$

4. Evaluation Metric

We will use the following two metrics: normalized mutual information (NMI)(Cai et al., 2011) and clustering accuracy (ACC)(Cai et al., 2011). All the above two measuring metrics are commonly used in the clustering literature, and all have pros and cons. Using them together will demonstrate total effectiveness of the clustering algorithm.

5. Future Work

We will implement IDEC(Xifeng Guo, 2017), Towards K mean(Yang et al., 2016) , VaDE(Jiang et al., 2016) and then compare the observed results using mentioned evaluation metrics.

References

Bengio, Yoshua, Lamblin, Pascal, Popovici, Dan, and Larochelle, Hugo. Greedy layer-wise training of deep networks. In Schölkopf, B., Platt, J. C., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems 19*, pp. 153–160. MIT Press, 2007.

Cai, Deng, He, Xiaofei, and Han, Jiawei. Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 23 (6):902–913, 2011.

Jiang, Zhuxi, Zheng, Yin, Tan, Huachun, Tang, Bangsheng, and Zhou, Hanning. Variational deep embedding: A generative approach to clustering. *CoRR*, abs/1611.05148, 2016. URL http://arxiv.org/abs/1611.05148.

Kingma, Diederik P. and Welling, Max. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL http://arxiv.org/abs/1312.6114.

Xie, Junyuan, Girshick, Ross B., and Farhadi, Ali. Unsupervised deep embedding for clustering analysis. *CoRR*, abs/1511.06335, 2016. URL http://arxiv.org/abs/1511.06335.

Xifeng Guo, Long Gao, Xinwang Liu Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 1753–1759, 2017. doi: 10.24963/ijcai.2017/243. URL https://doi.org/10.24963/ijcai.2017/243.

Yang, Bo, Fu, Xiao, Sidiropoulos, Nicholas D., and Hong, Mingyi. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. *CoRR*, abs/1610.04794, 2016. URL http://arxiv.org/abs/1610.04794.