DEEP CLUSTERING AND MIXED DATA: A REVIEW

Shabir Ahmad Rather
Research Scholar
Department of Computer Science
GDC Anantnag

Abstract: One of the basic problems in many data recovery applications is clustering. The performance of the pool depends on the representation of the data. In this article, we first briefly introduce deep clustering and big data. Then we mention the background and then we do a bibliographic survey. Finally, we present some future opportunities and conclusions in this field.

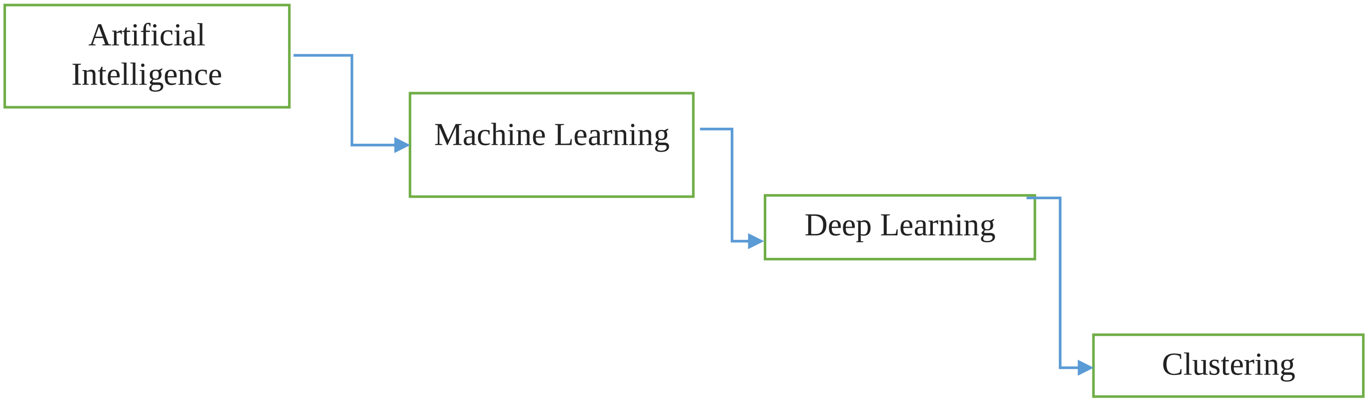
IndexTerms **- Deep Learning, Clustering, and Big data.**

I. Introduction

Today, deep learning and big data are two hot trends in the rapidly evolving digital world. Although big data has different definitions, it refers to exponential growth, which is difficult to manage and analyze using traditional software tools and techniques. Digital data comes in different forms and is growing at an alarming rate [1]. For example, according to the National Security Agency, the Internet processes 1,826 petabytes of data every day [2]. In 2011, the amount of digital information increased nine times in just five years [3], by 2020, its global total will reach 35 trillion GB [4]. Deep learning is one of the most important technologies in machine learning. It teaches computers what to do: learn by example. Most deep learning methods use neural architectures, which is why deep learning models are sometimes called deep neural networks. The term "depth" in neural networks generally refers to the number of hidden layers. Traditional neural networks only contain 23 hidden layers, while deep networks can have up to 150 hidden layers. The deep learning model uses tagged data for training and learns directly from functions without manually extracting them.

II. Background

The grouping is done on unstructured data, or we can say that it is heterogeneous and variable in nature, and has a variety of formats, including text, documents, images, videos, etc. According to a 2011 IDC study, it will represent 90% of all data created in the next ten years. The vast majority of newly produced substances are unstructured, which means that the information: 1) is heterogeneous (for example, log reports, videos, images, etc.), 2) does not have a standard model, and 3) comes from different sources [5]. Clustering [14] can be defined as the process of organizing objects into groups whose members are similar in some respects. Extracting relevant information from unstructured big data is time consuming and complex. The lack of structure makes compilation and analysis a time-consuming and labor-intensive task. When data sets are sorted or categorized by groups or categories, finding information from large data sets is easier and takes less time. The relationship between Deep Learning and Clustering is shown in the following figure:



The concept of “deep clustering” was first introduced in a deep learning framework to separate audio sources [6], and gradually became popular in general clustering tasks.

III. Literature Survey

The problem of clustering has been extensively studied in the database and statistics literature as part of numerous data mining tasks [7]. Similarity between objects is measured using a similarity function. Clustering is particularly useful for organizing documents, improving traceability and facilitating navigation [7]. The study of the clustering problem precedes its applicability to the textual domain. A broad overview of clustering (as it deals with general categorical and numeric data) can be found in [7]. Several implementations of popular text clustering algorithms, applied to textual data, can be found in some toolkits such as Lemur [8]. The clustering problem finds applicability for a number of tasks: document organization and navigation, corpus summary, document classification [7].

Discriminant clustering was first introduced by Xu et al. [9] and explicitly relies on supervised classification techniques such as support vector machines (SVMs) to perform unsupervised clustering: it aims to tag data so that if an SVM is executed with these tags, the resulting classification separates the data with a high margin. To solve the related association optimization problem on labels, Xu et al [9]. Consider the convex expansion in terms of a semi-defined program (SDP).

Bach and Hachioji [10], resulting in more suitable and feasible algorithms. Discriminative clustering is well suited to the segmentation problem for two reasons: first, we can reuse existing features for classification or supervised detection, especially architectures advanced based on local feature graphs and kernel methods [11]. Relying on previously researched and supervised tools specifically dedicated to fine-tuning these descriptors has been shown to be beneficial in other poorly supervised computer vision tasks [12, 13]. Second, discriminant clustering makes it easy to include constraints on the partitions found by the clustering algorithm, in our case local and spatial color consistency constraints.

Automatic speaker recognition is an important key technology in the path of machine learning semantic multimedia. It takes many forms: for example, speaker recognition refers to the task of inferring the identity of the speaker from a new utterance, given a set of known voice patterns. Speaker clustering describes the task of telling who spoke when for a sequence of utterances, without first knowing the number or identity of the speakers [14].

Zomaya et al. [15] presents a study of existing clustering algorithms of different categories (Partitioning, Hierarchical-Based, Density-Based, Grid-Based and Model-Based). In their work, they compared five categories with their most representative algorithm, their goal being to find the best performance for Big Data.

In [16] the authors focus on the most popular algorithms and the most used in the literature such as kmeans, they present some comparative works of these algorithms. Another recent study [17] provides a general overview of algorithms and data mining platforms that can be used in the field of big data by discussing together other challenges and characteristics.

Article [18] discusses several big data mining algorithms to find the most suitable of them using a full comparison.

In the context of deep learning for clustering, the two most dominant methods from each of these categories were used. Cluster clustering, which is a hierarchical clustering method, has been used with deep learning [19]. Every day, large amounts of data are generated from many sources. Thus, the term data is transformed into Big Data, facing challenges in information gathering and decision making. This processing of data can be aided by Deep Learning capabilities, in particular its ability to handle both labeled and unlabeled data that is often heavily collected in big data.

IV. FUTURE OPPORTUNITIES AND CONCLUSIONS

Since the deep cluster presents this feature extraction function, the clustering algorithm combines a deep learning for better consolidating results. In this article, we examine profound clustering and have popular research in grouping. It is very difficult to make a clustering with the help of traditional technologies because of the high complexity and calculation costs. As a result, the compactness and separation of data are one of the most important questions of the quality of the grouping. In the future, I would like to work for a deep flux for medical image analysis.

V. References

1. LIN, X.-W. C. (2014). Big Data Deep Learning: Challenges and Perspective. IEEEV ACCESS.
2. National Security Agency. (2013, August 9th). Retrieved from The National Security Agency: Missions,Authorities,OversightandPartnerships:http://www.nsa.gov/public_info/_files/speeches_testimonies/2013_08_09_the_nsa_story.pdf
3. Reinsel, J. G. (2011). Extracting Value from Chaos. Hopkinton, MA. USA: EMC,.
4. Reinsel, J. G. (2010). The Digital Universe Decade—Are You Ready. Hopkinton, MA, USA: EMC.

5. Mr. Anuj Kumar Ray, S. A. (2016). Unstructured Data and Term Mining based on Clustering in Document NOSQL. International Advanced Research Journal in Science, Engineering and Technology , Vol. 3, Issue 4.
6. J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, ''Deep clustering: Discriminative embeddings for segmentation and separation,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 31–35.
7. Aggarwal, C. Z. (2012). A survey of text clustering algorithms. In: Mining text data. Springer , 77-128.
8. Y. Zhou, H. C. (2009). Graph Clustering based on Structural/ Attribute Similarities. VLDB Conference.
9. L. Xu, J. N. (2005). Maximum margin clustering. NIPS.
10. Diffrac, F. a. (2007). a discriminative and fexible framework for clustering. NIPS.
11. J. Zhang, M. M. (2007). Local features and kernels for classifcation of texture and object categories: A comprehensive study. International Journal of Computer Vision, 213-238.
12. O. Duchenne, I. L. (2009). Automatic annotation of human actions in video. ICCV.
13. M. H. Nguyen, L. T. (2009). Rother Weakly supervised discriminative localization and classifcation: a joint learning process. ICCV.
14. Beigi, H. (2011). Fundamentals of speaker recognition. Springer Science & Business Media.
15. A. Fahad, N. A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis. IEEE transactions on emerging topics in computing.
16. A.BEN AYED, M. H. (2014). Survey on clustering methods: Towards fuzzy clustering for Big Data," In Soft Computing and Pattern Recognition (SoCPaR). 6th International Conference of. IEEE, 331-336.
17. A. Sherin, S. U. (2014). Survey On Big Data Mining Platforms, Algorithms And Challenges. International Journal of Computer Science & Engineering Technology.
18. S.ARORA, I. (2014). A survey of clustering techniques for Big Data analysis. 5th International Conference IEEE, 59-65.
19. Yang, J. P. (2016). Joint unsupervised learning of deep representations. Proceedings of the IEEE Conference on Computer Vision and Pattern, 5147–5156.
20. Hsu, C.-C. a.-W. (2017). Cnn-based joint clustering and representation learning with. Retrieved from arXiv preprint arXiv:1705.07091
21. Li, F. Q. (2017). Discriminatively boosted image clustering with fully. Retrieved from arXiv preprint arXiv:1703.07980.