Clustering Analysis

E. Goetz

I. Introduction

With the normalization of artificial intelligence, there is an ever growing hunger for data. Conglomerates such as Google, Facebook, and Amazon have the consumer base and power to mine and classify customer data, creating powerful training sets. But even these giants cannot collect enough information. Despite the amount of interest and wealth being poured into AI, no virtual machine has come close to general intelligence; many of the world's most powerful and advanced AI are incapable of learning without human intervention, effectively bottle-necking their abilities. But with the development of AlphaZero (the successor to AlphaGo and AlphaGo Master), the power of alternative models of machine learning, like reinforcement, are becoming ever more apparent [1]. Unfortunately, some tasks evade clear feedback and therefore evade algorithmic reinforcement. Perhaps the most notable of such tasks is one which humans do not traditionally associate with intelligence at all: perception. Specifically, the task of object, shape and pattern recognition. For tasks such as these, AI research has turned to unsupervised learning in the hope that one day machines will step in our shoes and see through our eyes. In pursuit of this dream, computer scientists have developed several algorithms and have carved out a unique section of machine learning: clustering.

II. Proposal

Many algorithms have been developed in order to cluster data, none of which are spoken of as being clearly superior to their rivals [2]. In response to the resulting wide and diverse field of approaches to clustering, this project will be an attempt to implement and analyze several common clustering algorithms and to see how their clusters differ from one another. Completing this analysis should reveal tentative data suggesting the strengths and weaknesses of different approaches.

Given their popularity, the following algorithms will be analyzed: k-means, agglomerative hierarchical clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Should time permit, other algorithms such as divisive hierarchical clustering, soft k-means, and spectral clustering may be included. Any included algorithms will be naively implemented in Python with the intent of developing their efficiency in the future. The module scikit-learn will then be used to generate artificial data sets and to test the functionality of the implemented algorithms. No specialized hardware will be used.

Once having obtained the clustering results, the various algorithm's clusters will be analyzed using the S_Dbw index. S_Dbw is designed to score clustering results based on "clusters' compactness (in terms of intra-cluster variance)" as well as "density between clusters (in terms of inter-cluster density)" [3]. While there are many indices, this particular index has been chosen based on previous performance evaluations [4], [5].

$$S\_Dbw(c) = Scat(c) + Dens\_bw(c)$$

Fig. 1. The S_Dbw index formula. $Scat(c)$ represents the average scattering for clusters and $Dens\_bw(c)$ represents inter-cluster density.

III. Conclusion

By analyzing the abilities of various algorithms, this project will provide a coherent introduction to clustering and illuminate the progress in unsupervised learning. With further research, this project may indicate why certain algorithms succeed in analyzing some clusters and yet fail tragically in analyzing others. Obtaining such information will allow for the modification and— potentially— for the discovery of algorithms. Resultantly, clustering technique analysis is an indispensable step in developing the next generation of unsupervised intelligence.

References

[1] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, vol. 550, October 2017. [Online]. Available: https://doi.org/10.1038/nature24270

[2] M. K. Paisarn, M. Kambiz, and J. L. Guan, *Unsupervised learning*, 1st ed. John Wiley & Sons, Inc., May 2014. [Online]. Available: https://preview.tinyurl.com/ybbdy5ae

[3] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a data set," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 02 2001, pp. 187–194. [Online]. Available: https://preview.tinyurl.com/yaxn9dmk

[4] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *2010 IEEE International Conference on Data Mining*, 2010. [Online]. Available: https://preview.tinyurl.com/y7urk4zn

[5] C. Legany, S. Juhasz, and A. Babos, "Cluster validity measurement techniques," in *WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases*, February 2006. [Online]. Available: https://preview.tinyurl.com/y8nf3p9e