A NEW PERSPECTIVE TO OVERCOME THE CHALLENGE OF SEMI-SUPERVISED CLUSTERING PROBLEM IN MACHINE LEARNING

Ali Hakan Tor[1,*]

[1]*Department of Engineering Science, Faculty of Engineering, Abdullah Gül University, Kayseri, Türkiye*

ABSTRACT

Machine learning (ML), a subfield of artificial intelligence (AI), can be defined as "learning" methods. In other words, they are methods developed without being explicitly programmed to make predictions and decisions using available data. The study of mathematical optimization offers methods, theory, and application areas for machine learning. For instance, speech recognition [10, 12, 8], internet security [7, 5], road detection [3, 1], image classification [2, 11, 6], information retrieval [4], bioinformatics [9], and many more. These are examples of machine learning for supervised, unsupervised, semisupervised, self-supervised, and reinforcement learning. These are examples of machine learning for supervised, unsupervised, semisupervised, self-supervised, and reinforcement learning. This study focuses on semisupervised learning problems as in [10, 12, 7, 3, 2, 4, 9], in particularly semisupervised clustering. When must-link and cannot-link constraints are given for a clustering problem, this clustering problem is considered a semi-supervised learning problem. Must-link constraints force certain data points to be in the same cluster, while cannot-link constraints prevent some data points from being in the same cluster, which can create complex requirements for the clustering algorithm to balance. Ensuring that these constraints are satisfied while forming coherent and meaningful clusters can be difficult, especially when the constraints are sparse or conflict with the natural structure of the data. Especially when the number of clusters increases, the rate of providing must-link constraints decreases dramatically. Another difficulty is the computational complexity involved in integrating these constraints into the clustering process, particularly for large datasets. The algorithm must efficiently search for clusters that satisfy the given constraints while maintaining overall cluster quality. Must-link constraints form an equivalence relation. In this context, it is possible to separate must-link constraints into equivalence classes. In this presentation, a suggestion based on this equivalence class will be given to satisfy more must-links for the new algorithms.

Keywords Machine Learning · Semi-supervised Clustering · Non-smooth Optimization · Semi-supervised Learning

References

[1] DOLLAR, P., TU, Z., AND BELONGIE, S. Supervised learning of edges and object boundaries. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (2006), vol. 2, IEEE, pp. 1964–1971.

[2] GONG, C., TAO, D., MAYBANK, S. J., LIU, W., KANG, G., AND YANG, J. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing 25*, 7 (2016), 3249–3260.

[3] HAN, X., LU, J., ZHAO, C., YOU, S., AND LI, H. Semisupervised and weakly supervised road detection based on generative adversarial networks. *IEEE Signal Processing Letters 25*, 4 (2018), 551–555.

[*] *Corresponding Author's E-mail: hakan.tor@agu.edu.tr*

[4] Hussain, A., and Cambria, E. Semi-supervised learning for big social data analysis. *Neurocomputing 275* (2018), 1662–1673.

[5] Ioannou, C., and Vassiliou, V. Classifying security attacks in iot networks using supervised learning. In *2019 15th International conference on distributed computing in sensor systems (DCOSS)* (2019), IEEE, pp. 652–658.

[6] Ma, L., Li, M., Ma, X., Cheng, L., Du, P., and Liu, Y. A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing 130* (2017), 277–293.

[7] Rathore, S., and Park, J. H. Semi-supervised learning based distributed attack detection framework for iot. *Applied Soft Computing 72* (2018), 79–89.

[8] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), IEEE, pp. 6989–6993.

[9] Roder, H., Oliveira, C., Net, L., Linstid, B., Tsypin, M., and Roder, J. Robust identification of molecular phenotypes using semi-supervised learning. *BMC bioinformatics 20*, 1 (2019), 1–25.

[10] Sholokhov, A., Sahidullah, M., and Kinnunen, T. Semi-supervised speech activity detection with an application to automatic speaker verification. *Computer Speech & Language 47* (2018), 132–156.

[11] Tuia, D., Volpi, M., Copa, L., Kanevski, M., and Munoz-Mari, J. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing 5*, 3 (2011), 606–617.

[12] Yu, D., Varadarajan, B., Deng, L., and Acero, A. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. *Computer Speech & Language 24*, 3 (2010), 433–444.