IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 17, 2021, accepted August 7, 2021, date of publication August 12, 2021, date of current version August 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104357

A Deep-Learned Embedding Technique for Categorical Features Encoding

MWAMBA KASONGO DAHOUDA AND INWHEE JOE
Department of Computer Science, Hanyang University, Seoul 04763, South Korea

Corresponding author: Inwhee Joe (iwjoe@hanyang.ac.kr)

This work was supported by the Institute for Information and Communication Technology Promotion (IITP) funded by Korea Government [Ministry of Science, ICT and Future Planning (MSIP)] (Development of the technology to automate the recommendations for big data analytic models that define data characteristics and problems) under Grant 2020-0-00107.

ABSTRACT Many machine learning algorithms and almost all deep learning architectures are incapable of processing plain texts in their raw form. This means that their input to the algorithms must be numerical in order to solve classification or regression problems. Hence, it is necessary to encode these categorical variables into numerical values using encoding techniques. Categorical features are common and often of high cardinality. One-hot encoding in such circumstances leads to very high dimensional vector represen-