Deep Clustering for Mixed-type Data with Frequency Encoding and Doubly Weighted Cross Entropy Loss

Deogho Choi
Department of Semiconductor and Display Engineering
Sungkyunkwan University
Suwon, Republic of Korea
sailoer123@skku.edu

Daniel Chae
Infra OSS Tech. Lab
SK Telecom
Seoul, Republic of Korea
dani75.chae@sk.com

Wooyeon Kim
Infra OSS Tech. Lab
SK Telecom
Seoul, Republic of Korea
wooyeon.kim@sk.com

Jihong Kim
Infra OSS Tech. Lab
SK Telecom
Seoul, Republic of Korea
mittens@sk.com

Janghoon Yang*
Department of AI Software Engineering
Seoul Media Institute of Technology
Seoul, Republic of Korea
jhyang@smit.ac.kr

Jitae Shin*
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
jtshin@skku.edu

***Abstract*—Clustering algorithm is unsupervised learning that groups a set of data into distinctive classes according to the similarity between each data sample. Most of previous researches have focused on improving K-prototypes or training proper numerical representations of categorical features using autoencoder. But in this research, we investigate that applying frequency encoding to categorical features can be sufficiently effective. Furthermore, we propose doubly weighted cross entropy loss, DW-CE loss, to find optimal cluster centroid by training fully connected layer. The experiment with two mixed-type datasets, credit approval and heart disease, from UCI repository shows that the proposed clustering with frequency encoding and DW-CE loss provides better performance than existing state of the arts methods in most of cases.**

Keywords—clustering, mixed-type, neural network, nonnumeric data encoding, autoencoder

I. INTRODUCTION

Clustering algorithm groups a set of unlabeled data into distinctive classes according to the similarity between each data sample. Depending on the purpose of clustering, clustering can be different for the same dataset. Thus, the characteristic of each grouped class can be used as a useful information to provide a target service. Many conventional clustering algorithms, such as k-means [1], DBSCAN [2], and spectral clustering [3], are designed for numerical features from which we can calculate the distance between each category. But most real-world datasets are mixed-type that includes both numerical and categorical features. Some categorical features like disease severity have a clear order, and others like sex don't have a clear order. But in both cases, we don't know the exact distance between each category. Due to the ambiguity of measuring the distance of categorical features, it is difficult to apply conventional clustering algorithms directly in mixed-type dataset.

One common approach to solve this problem is treating numerical and categorical features separately. K-prototype [4] is a basic algorithm using this concept. This algorithm combines k-means for numerical features and k-modes for categorical features. In this case, Hamming distance can be applied to one-hot encoded categorical features for calculating the distance [5][6]. But Hamming distance only indicates the existence of a mismatch by value 1. In this case, all categorical features are placed at equal distances, and thus sophisticated distances according to feature values cannot be considered.

To avoid this problem, the autoencoder can be used for training proper numerical representations of categorical features to apply conventional clustering algorithms. A trained latent vector has the advantage of compactness because one-hot encoded vector can be long if the categorical feature has various values, and it also has the advantage of being able to learn latent vectors tailored to the appropriate purpose with an additional loss function [7][8]. However, rare categorical values can be omitted due to the data imbalance in the process of training with reconstruction loss.

In this paper, we investigate that applying frequency encoding without training autoencoder to learn proper numerical representations of categorical features can be effective with conventional clustering algorithms. In addition, although most existing papers use conventional machine learning methods as core clustering algorithms, we propose doubly weighted cross entropy loss (DW-CE loss) to find optimal cluster centroids by updating weights of fully-connected layer.

II. DEEP CLUSTERING

The framework of our proposed deep clustering is described in Figure 1. It consists of two main parts: preprocessing and model training. In preprocessing part, categorical features are converted into numerical features through frequency encoding. With preprocessed inputs, clustering model is trained such that the weights of fully

978-1-6654-8559-3/22/$31.00 ©2022 IEEE

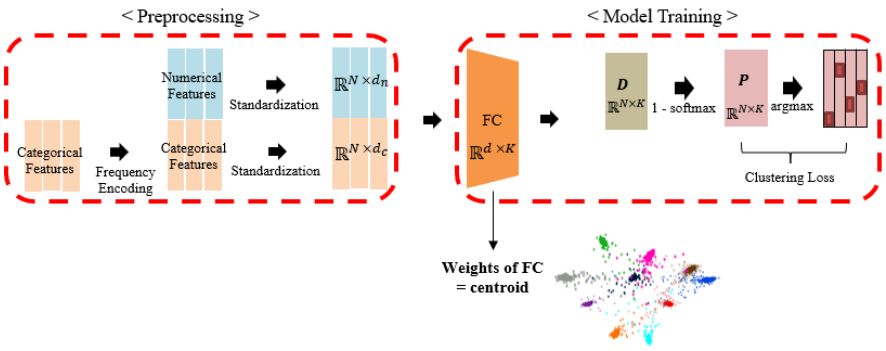

Fig. 1. Framework of deep clustering

connected layer can work as cluster centroid with articulated clustering loss.

A. *Frequency encoding*

Domain expert may be able to decide which features are particularly important. However, if such knowledge is not present, it is inevitable to assume that all features have the same weight. In this case, if there is a particularly large difference in one feature, the influence of that feature is more dominant than that of the other in the clustering process. However, if Hamming distance which is commonly used for calculating the distance of categorical features, is applied, all possible distances are 0 and 1. As a result, the distance diversity of the categorical features is insufficient compared to numerical features, and the priority of categorical features could be relatively low. To tackle this limitation, we can apply frequency encoding to categorical features. In the case of frequency encoded features, if two values appear at similar frequencies, the distance between these two values is measured small, and if one value appears frequently and the other value appears occasionally, the distance is measured large. So, if the frequency of two categorical values is very different, the distance can be measured large, and it imposes high priority of that categorical feature in the clustering process. Frequency encoding is described in Figure 2, each categorical feature is encoded with the frequency of an occurrence throughout entire dataset.

Height	Sex
173.1	Male
160.4	Female
178.5	Male
155.5	Female
163.7	Female

Frequency Encoding

Height	Sex
173.1	0.4
160.4	0.6
178.5	0.4
155.5	0.6
163.7	0.6

Fig. 2. Example of frequency encoding

B. *Doubly weighted cross entropy loss*

Let $X = \{x_1, x_2, \dots, x_N\}$ be a set of N samples in the given dataset, which each data sample x_i has d features consist of d_c categorical features and d_n numerical features. Numerical features can be denoted as $F_n = \{f_n^1, \dots, f_n^{d_n}\}$ and categorical features can be denoted as $F_c = \{f_c^1, \dots, f_c^{d_c}\}$. This categorical feature f_c represents frequency encoded categorical feature. In this paper, frequency encoded categorical feature f_c can be treated in the same way as numerical feature f_n for clustering. After frequency encoding, standardization method is applied to both numerical feature f_n and frequency encoded categorical feature f_c. Inspired by [9][10], we set the weights $W \in R^{d\times k}$ of fully connected layer as centroids $\mu \in R^d$ of k clusters. These weights are initialized using standardization method. The Minkowski distance, which is defined in (2) is applied to calculate distance $D_{i,k}$ between each data sample x_i and each cluster centroids μ_k, and then the nearest cluster z among k clusters is used as pseudo label. Following the equation (3), the probability $p_{i,k}$ of being assigned to each cluster can be calculated by applying the softmax function to the distances $D_{i,k}$, and then subtracted from 1. This probability $p_{i,k}$ can be used to calculate cross-entropy loss with nearest cluster z as pseudo label. However, if we only use this cross-entropy loss for updating cluster centroids, the training process can be unstable due to outlier data samples and cluster imbalance problem. To stabilize training process, two hyper-parameters, sample weight and cluster weight, can be exploited for defining a clustering loss. Sample weights$_i$imposes influence on each data sample depending on probability $p_{i,z}$. It means data sample x_i which has low probability $p_{i,z}$ can be considered as outlier, so the update of centroid μ_z by data sample x_i is less affected compared to other samples. Cluster weight c_i is needed to escape cluster imbalance problem. Since clustering is unsupervised learning that does not use labels during training, the process that find optimal cluster centroids using loss function would lead to trivial solution where the clusters are collapsed into single entity. Cluster weight imposes influence on each data sample depending on cluster frequency. It means that if the nearest cluster z of data sample x_i has large number of assigned data samples, the loss value of x_i has low influence. It derives the degree of influence depending on cluster size, and it gives opportunity to assign more data sample to relatively small cluster.

$$\text{DW-CE } loss = -\frac{1}{N}\sum_{i=1}^{N} s_i \times c_i \times \log p_{i,z} \quad (1)$$

$$d_{i,k} = \|x_i - \mu_k\|_p \quad (2)$$

$$p_{i,k} = 1 - \frac{d_{i,k}}{\sum_{j=1}^{k} e^{d_{i,j}}} \quad (3)$$

$$s_i = \max_k p_{i,k} = p_{i,z} \quad (4)$$

$$c_i = 1 - \frac{|Cluster_z|}{\sum_{j=1}^{k}|Cluster_j|} \quad (5)$$

III. EXPERIMENTS

A. Datasets

In this research, two mixed-type datasets were used from UCI repository, which are heart disease and credit approval dataset. The brief description about these datasets is shown in Table 1. The missing values were imputed by mean value. To realign all features, standardization method was applied with (6).

$$z = \frac{x-\mu}{\sigma} \tag{6}$$

TABLE I. DESCRIPTION OF THE DATASETS

	samples	d_n	d_c	***missing***	***class***
Credit	690	6	8	37	2
Heart disease	303	6	7	0	2

B. Evaluation metrics

In this research, we evaluate the performance of the proposed deep clustering with four evaluation metrics.

Purity evaluates the frequency of true positives assuming that predictions are assigned to most common ground truth class.

$$Purity(U,V) = \frac{1}{N}\sum_k \max_j \left|U_k \cap V_j\right| \tag{7}$$

Rand index (RI) is a function that measures the similarity of the two assignments. Rand index can range from 0 to 1. Lower value indicates different labeling, while higher value indicates similar labeling between ground truths and predictions.

$$RI = \frac{TP+TN}{TP+FP+FN+TN} \tag{8}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative with ground truth.

Normalized mutual information (NMI) is a normalization of the Mutual Information (MI) score to account for agreement of the two assignments.

$$NMI(U,V) = \frac{\sum_k \sum_j P(U_k \cap V_j) \log \frac{P(U_k \cap V_j)}{P(U_k)P(V_j)}}{\sqrt{(\sum_k P(U_k) \log P(U_k)) \times (\sum_j P(V_j) \log P(V_j))}} \tag{9}$$

Folkes and Mallows index (FMI)is defined as the geometric mean of the pairwise precision and recall.

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{10}$$

C. Clustering results

The proposed deep clustering with one fully connected layer was implemented with Pytorch package. Adam optimizer was used with learning rate 0.1. The order of the Minkowski distance to calculate distance between each data sample and cluster centroid was heuristically chosen as 3 on credit approval dataset case and 2 on heart disease dataset case. The missing values were imputed by mean value. To assess the effect of the categorical features on clustering, K-prototype with all features and K-means with numerical features only was compared. The K-prototypes was shown to provide marginally better performance with the additional information of categorical features compared to K-means with numerical features only. But K-means algorithm with numeric features and frequency-encoded categorical features was found to increase the metrics significantly in comparison to the K-prototypes, which proves the efficacy of the frequency encoding. The proposed clustering with frequency encoding and a doubly weighted entropy loss is found to achieve the highest value for all considered metrics over credit approval dataset while it does for Purity and FMI over heart disease dataset, which verifies the superiority of the proposed method.

TABLE II. PERFORMANCE ON CREDIT APPROVAL DATASET

	Purity	***RI***	***NMI***	***FMI***
K-means (Only numerical)	0.64	0.54	0.08	0.63
K-prototypes	0.67	0.56	0.13	0.68
K-means (Frequency encoded)	0.81	0.70	0.32	0.71
Multi-view K-prototypes [5]	0.81	0.70	-	-
GA-FKP [6]	0.86	-	-	-
COPE [7]	-	-	-	0.77
KMFM [8]	-	0.70	0.34	-
UFL Fuzzy ART [10]	0.86	0.75	-	-
AMDPC [11]	0.83	0.72	0.35	-
Proposed method	**0.87**	**0.77**	**0.46**	**0.77**

TABLE III. PERFORMANCE ON HEART DISEASE DATASET

	Purity	***RI***	***NMI***	***FMI***
K-means (Only numerical)	0.71	0.58	0.12	0.59
K-prototypes	0.76	0.63	0.19	0.63
K-means (Frequency encoded)	0.82	0.71	0.33	0.71
Multi-view K-prototypes [5]	0.81	0.69	-	-
GA-FKP [6]	-	-	-	-
COPE [7]	-	-	-	-
KMFM [8]	-	**0.72**	**0.35**	-
UFL Fuzzy ART [10]	0.82	0.70	-	-
AMDPC [11]	-	-	-	-
Proposed method	**0.83**	0.71	0.34	**0.72**

IV. CONCLUSION

We investigate the effectiveness of frequency encoding in the absence of expert knowledge. It has the advantage in terms of performance, and it also has the simple structure providing computational advantage over other previous researches that entail auto-encoder. Furthermore, the proposed clustering loss is found to contribute to providing better performance in most of the cases than other previous researches. There are still several limitations to be studied. We used fully connected layer to find optimal cluster centroid. A powerful function approximation which can be achieved by deep layers and articulated structures need to be studied to improve clustering performance further.

Clustering results are found to have significant dependency on initiations. Thus, articulated initialization is likely to improve the clustering performance further. Also, in this research we used frequency encoding to find numerical representation of categorical feature. But it needs supervision and assumes static situation. The way to apply frequency encoding to categorical features in online learning situation will be studied in the future.

ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)(2020-0-01305, Development of AI Deep-Learning Processor and Module for 2,000 TFLOPS Server)

This research was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01798) supervised by the IITP(Institute for Information & communications Technology Promotion)

REFERENCES

[1] A. K. Jain and R. C. Dubes, "Algorithms for clustering data", *Technometrics*, ACM, pp. 227-229, 1988.

[2] M. Ester, et al, "A density-based algorithm for discovering clusters in large spatial databases with noise", *KDD 96*, ACM, pp. 226-231, 1996.

[3] U. von Luxburg, "A tutorial on spectral clustering", *Statistics and Computing*, Springer, pp.395-416, 2007.

[4] Z. Huang, "Clustering large data sets with mixed numeric and categorical values", *PAKDD*, Springer, pp. 21-34, 1997.

[5] J. Jinchao, et al, "A Multi View Clustering Algorithm for Mixed Numeric and Categorical Data", *IEEE Access*, IEEE, pp. 24913-24924, 2021.

[6] R. Nooraeni, M. I. Arsa, and N. W. K. Projo, "Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data", *Procedia Computer Science*, Elsevier, pp. 677-684, 2021.

[7] L. Tran, L. Fan, and C. Shahabi, "Clustering Mixed-Type Data with Correlation-Preserving Embedding", *International Conference on Database Systems for Advanced Applications*, Springer, pp. 342-358, 2021.

[8] S. Sahoo, and S. Chakraborty, "Learning Representation for Mixed Data Types with a Nonlinear Deep Encoder-Decoder Framework", *arXiv preprint*, 2009.09634, 2020.

[9] K. Kenyon-Dean et al, "Clustering-Oriented Representation Learning with Attractive-Repulsive Loss", *arXiv preprint*, 1812.07627, 2018.

[10] D. Lam, M. Wei, and D. Wunsch, "Clustering data of mixed categorical and numerical type with unsupervised feature learning", *IEEE Access*, IEEE, pp. 1605-1613, 2015.

[11] L. Shihua, "Adaptive Mixed-Attribute Data Clustering Method Based on Density Peaks", *Complexity*, Hindawi, pp. 1-13, 2022.