Preprocessing categorical data with embeddings for K-Means

In this work, we explored the efficacy of utilizing embeddings for data preprocessing in k-means clustering, aiming to pinpoint specific dataset characteristics that determine when this innovative approach outperforms traditional methods to enhance the precision and reliability of clustering outcomes.

Background

The focus of our study was sparked by an article [1] that introduced the idea of using embeddings from large language models (LLMs) for k-means clustering with categorical data. This method, it suggested, could potentially be more effective than traditional techniques like one-hot encoding, power transformers, and ordinal encoders. The rationale is that embeddings can capture a more complex and nuanced representation of categorical variables, leading to improved clustering results. Inspired by this, our work aims to put this theory to the test. We also applied this embedding-based approach to various datasets. Our primary objective is to uncover the characteristics of datasets for which this innovative method demonstrates superior performance and can be most valuable.

Methodology

We selected three datasets to assess embeddings for k-means clustering on categorical data, conducting extensive exploratory data analysis (EDA) to understand their characteristics and variance. Following this, we applied one-hot encoding (OHE) to categorical variables, power transformer to numerical, and ordinal transformer to ordinal categorical data before performing k-means clustering. We evaluated the results using silhouette and Davies-Bouldin scores. Subsequently, we encoded the datasets with embeddings from a sentence transformer and conducted clustering again, measuring the performance with the same metrics. This approach allowed for a direct comparison between traditional encoding methods and embedding-based techniques.

Results

The clustering with embeddings got better for two of these datasets – both the Davies-Bouldin and silhouette scores showed this improvement. These two datasets had a similar amount of variance in their samples. However, the third dataset has shown different results. When we used embeddings on this one, the clustering results actually got worse, as shown by lower Davies-Bouldin and silhouette scores. This dataset had a lot more sample variance – almost double compared to the first dataset and almost three times more than the second. This shows that the amount of variance in a dataset can affect how well the embeddings work for k-means clustering. It's a key factor we must consider when choosing this clustering method. The actual results are presented on figure 1.

Results of Non-Embedded Data

Dataset	*Sample Variance*	*Mean Categorical Data Variance*	*Within Cluster Variance*	*Silhouette Score*	*Davies-Bouldin Index*
Customer Segmentation	0.08	0.02	0.08	0.20	1.85
Student Exam Performance	0.17	0.04	0.02	0.19	1.49
Bike Buyers	0.06	0.04	0.02	0.19	1.64

Results of Embedded Data

Dataset	*Silhouette Score*	*Davies-Bouldin Index*
Customer Segmentation	0.33	1.19
Customer Segmentation adj.	0.33	1.23
Student Exam Performance	0.22	1.71
Bike Buyers	0.27	1.45

Figure 1

Future work
Future projects might evaluate the same data encoding methodology with different machine learning techniques, such as linear regression. Another possible future project is to find a threshold in the sample variance in data, after which the implication of embeddings will not bring any extra value.

References

1. https://towardsdatascience.com/mastering-customer-segmentation-with-llm-3d9008235f41#3a33
2. https://github.com/moshchev/data_analytics_ws23