*Advances in Technology Innovation, vol. 7, no. 1, 2022, pp. 30-40*

# Clustering Analysis with Embedding Vectors: An Application to Real Estate Market Delineation

Changro Lee*

Department of Real Estate, Kangwon National University, Chuncheon, South Korea

Received 16 September 2021; received in revised form 15 November 2021; accepted 16 November 2021

DOI: https://doi.org/10.46604/aiti.2021.8492

## Abstract

Although clustering analysis is a popular tool in unsupervised learning, it is inefficient for the datasets dominated by categorical variables, e.g., real estate datasets. To apply clustering analysis to real estate datasets, this study proposes an entity embedding approach that transforms categorical variables into vector representations. Three variants of a clustering algorithm, i.e., the clustering based on the traditional Euclidean distance, the Gower distance, and the embedding vectors, are applied to the land sales records to delineate the real estate market in Gwacheon-si, Gyeonggi province, South Korea. Then, the relevance of the resultant submarkets is evaluated using the root mean squared errors (RMSE) obtained from a hedonic pricing model. The results show that the RMSE in the embedding vector-based algorithm decreases substantially from 0.076-0.077 to 0.069. This study shows that the clustering algorithm empowered by embedding vectors outperforms the conventional algorithms, thereby enhancing the relevance of the delineated submarkets.

**Keywords:** clustering, categorical data, high-cardinality, entity embedding, market delineation

## 1. Introduction

Machine learning is rapidly expanding to various applications; particularly, it has been used with great success in several applications such as computer vision, natural language processing, speech recognition, and time-series forecasting [1-4]. There are two main types of tasks within the field of machine learning: supervised and unsupervised learning. In a supervised learning framework, the algorithm learns on a labeled dataset. However, an unsupervised learning framework provides unlabeled data that the algorithm attempts to learn by extracting meaningful features without any guidance from the labels. Clustering is a de facto standard tool employed in unsupervised learning and is extensively used as a technique for discovering hidden patterns in data, such as consumer groups based on demographic profiles, or real estate submarkets based on property characteristics [5].

Although clustering analysis is a popular tool in unsupervised learning, it experiences difficulty when categorical variables are dominant in the dataset. Clustering is the process of grouping similar data points. The similarity between data points is calculated by a distance measure, which is essentially based on the geometry and distance in the Euclidean space. This concept of physical distance is well-suited to continuous data, but it is not directly applicable to categorical data. For instance, categorical data such as the color of products with each element being black, blue, and red cannot be clustered based on the distance between the three colors. Several methods, such as one-hot encoding approaches and specialized distance metrics for categorical data, have been proposed in the literature to solve this problem; however, these methods have not performed satisfactorily. This study attempts to overcome this limitation by using an entity embedding approach. The entity embedding maps categorical data into metric spaces, such as the Euclidean space, and thus can alleviate the discrete properties inherent in categorical data.

* Corresponding author. E-mail address: spatialstat@naver.com

Tel.: +82-33-250-6833

This study offers a way to enhance the clustering performance via an entity embedding approach, which has been actively used in the field of machine learning, particularly in natural language processing tasks. First, a study area is chosen to delineate the real estate market. Because the real estate market is always localized, it is routine to segment the market before performing the main tasks, such as property price estimation for tax assessment [6]. Second, to construct appropriate submarkets within the study area, the sales records of lots sold from 2016 to 2018 are analyzed and grouped by using several variants of a *k*-means clustering algorithm, respectively. An entity embedding approach is applied to handle high-cardinality variables when performing clustering analysis. Finally, the predictive accuracy of different sets of submarkets created earlier is compared in the context of property valuation.

The contribution of this study can be summarized as follows:

(1) By adopting the framework proposed in this study, clustering algorithms can be utilized in a more efficient manner when applied to the tasks in which high-cardinality categorical data are dominant.

(2) This study shows that the entity embedding approach can be used effectively in processing structured data, i.e., traditional tabular format data with rows and columns, beyond the field of unstructured data, such as images and free-form texts.

(3) To the best of the authors' knowledge, real estate markets have not been delineated with the aid of entity embedding. This study attempts to construct real estate submarkets by exploiting the entity embedding approach.

The remainder of this study is organized as follows. Section 2 presents the background information on clustering analysis and real estate market delineation. Section 3 describes the dataset, embedding vectors, and silhouette score used to determine the optimal number of clusters. The results and interpretations are provided in section 4. A summary of this study and conclusions are presented in section 5.

## 2. Literature Review

### *2.1. Clustering analysis and entity embedding*

For a clustering algorithm to group observations together, the notion of (dis)similarity among observations must be defined first. A popular dissimilarity metric, i.e., a distance metric for clustering, is the Euclidean distance. However, the Euclidean distance is only applicable to continuous variables and not categorical variables. When categorical variables must be used for clustering, the simplest technique to calculate the Euclidean distance is to convert each element in a categorical variable to a separate binary dummy variable, and this is often called the one-hot encoding approach. Although this approach is applicable to a clustering algorithm, it becomes cumbersome and inefficient for processing data when categorical variables are dominant in the dataset.

Alternatively, a special distance metric that can handle both continuous and categorical data types may be utilized, and a well-known such metric is the Gower distance [7]. To calculate the Gower distance, an appropriate distance metric for each variable type is selected and scaled to fall between 0 and 1; the Euclidean distance is chosen for continuous variables, and the Dice coefficient is chosen for categorical variables [8]. The Dice coefficient is a well-known similarity measure for categorical data [9]. Subsequently, the average value of all variables is calculated to create the final Gower distance. The Dice coefficient is a measure that replaces the Euclidean distance metric with a matching similarity measure, which is easily applicable to categorical data. Many clustering algorithms proposed for grouping categorical variables are implemented based on the modified concept of (dis)similarity, such as the Dice coefficient; K-modes is one such well-known algorithm [10]. K-modes is a clustering algorithm that is specialized in classifying categorical data. In addition, useful modifications to the K-modes have been proposed recently to achieve better performance [11-12].

However, the above approaches become very inefficient when a categorical variable is highly cardinal; that is, the number of unique values in a categorical variable is very large. To solve this problem, this study utilizes an entity embedding approach for high-cardinality categorical variables. An entity embedding technique is used for mapping categorical values to a multidimensional space with fewer dimensions than the original number of levels, where the values with similar function outputs are close to each other [13-14]. An entity embedding is usually created via neural network training in the form of embedding vectors. The entity embedding technique is widely used in the field of natural language processing because words can be viewed as the agglomeration of high-cardinality categorical variables [15-18].

According to the work of Guo et al. [13], the traditional one-hot encoding of a categorical variable can be expressed as:

$$x_i \to \delta_{x_i a} \tag{1}$$

where $\delta_{x_i a}$ denotes the Kronecker delta, and the possible values for $a$ are the same as $x_i$. The element is nonzero only when $a = x_i$. Based on the work of Guo et al. [13], the entity embedding of $x_i$ can be expressed as:

$$x_i \to \sum_a w_a \delta_{x_i a} = w_{x_i} \tag{2}$$

where $w_a$ represents the weight connecting the one-hot encoding layer to the embedding layer in a neural network. Thus, the mapped embedding is the weight of the embedding layer and can be learned in the same way as the parameters of other neural network layers.

In short, entity embedding is an approach of converting a categorical variable into a vector representation. The entity can be a word in natural language processing or a level in a categorical variable. Synonyms in natural language or similar levels in a categorical variable come to have similar vector values after neural network training. Fig. 1 illustrates the vector representation of entity embedding for the categorical variable "pet breeds".

A few studies utilizing entity embedding for clustering have been reported in the literature [19-20]. However, they attempted to apply the entity embedding technique to unstructured data such as imagery data, more specifically, the Modified National Institute of Standards and Technology (MNIST) database. This image dataset is already pre-processed public data. This study differs from them in that the entity embedding is applied to structured data, that is, tabular format data with rows and columns, and the real estate dataset used in the study is collected from scratch and cleaned thoroughly.

By adopting the entity embedding approach, especially for the data with high-cardinality categorical variables, several problems can be mitigated. First, an unrealistic amount of computational resource consumption can be avoided owing to the traditional one-hot encoding of high-cardinality variables. Second, different values of categorical variables can be treated in a meaningful manner, instead of processing them completely independently from each other. Third, the feature engineering step is avoided because the embeddings, by nature, can intrinsically group similar values together, thereby removing the need for domain experts to learn the relationships between values in the same categorical variable. Finally, the learned embedding vectors can be visualized using a dimensionality reduction technique, providing additional insights to business practitioners.

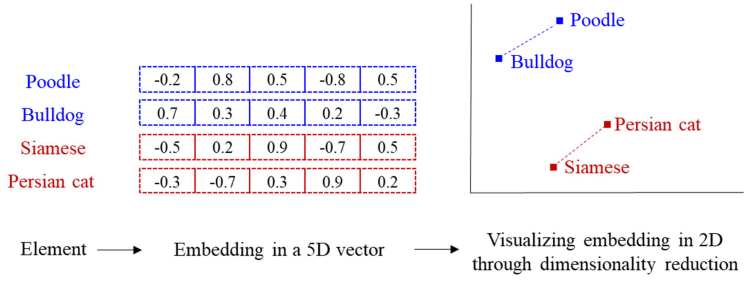



Fig. 1 Vector representation of entity embedding in the case of "pet breeds"

In this study, the embedding vectors are created by neural network training for high-cardinality variables and added to the existing dataset that comprises continuous variables. Subsequently, to overcome the inefficiency observed in the one-hot encoding and Gower distance approaches, the Euclidean distance is calculated for this combined dataset.

### *2.2. Application to real estate market delineation*

Because buyers and sellers require different types of properties for different reasons, the real estate market is almost always divided into submarkets within those property types and sales motivations. Because real estate is best characterized by its fixed location, its market delineation usually involves defining geographical boundaries [21-22]. It is routine that licensed appraisers and automated valuation models first identify the relevant submarket boundaries and then analyze the supply and demand of the proposed properties within a specific submarket. A submarket is generally associated with a group of similar properties in terms of price level and geographical location.

Various methods have been proposed to delineate the real estate market; examples of these methods range from well-established approaches, such as automatic zoning procedure [23-24] and spatial 'k'luster analysis by tree edge removal algorithm [25-26], to more recently suggested methods, such as adaptive density-based spatial clustering [27-28]. Most of these methods are based on a range of data-driven algorithms. The most popular algorithm for market delineation is a clustering algorithm. Conversely, a clustering procedure is performed using relevant variables such as price and geographical location, following which the clustering results are evaluated considering both internal and external validity measures [29-30]. Representative indices for internal validity measures include inertia and silhouette score; however, the criteria for external validity measures can vary depending on the dataset and research purpose [31-34].

This study adopts a *k*-means clustering algorithm to group similar observations (sales records of land in this study) into several distinct clusters, which serve as real estate submarkets for the purpose of price estimation. Several studies utilized price estimation results to test market segmentation, and they proved that market delineations improved the accuracy of price estimation [35-37].

## 3. Embedding Vectors and Silhouette Score

### *3.1. Dataset*

This study estimates the land prices in Gwacheon-si, Gyeonggi province, South Korea. With a population of over 13 million, Gyeonggi province is the most populous province in South Korea [38]. The Gwacheon-si government is one of the 45 local governments in Gyeonggi province, and Gwacheon-si comprises both urban and rural areas. This mixed landscape is the main reason behind choosing Gwacheon-si for the analysis because the heterogeneity strongly indicates the need for market segmentation.

The dataset is periodically provided in a comma-separated value file format by the Ministry of Land, Infrastructure and Transport (MOLIT) website, and have been released by the government since 2006, when the Act on Report on Real Estate Transactions was enforced. These land sales records are used in a range of government administrations, such as monitoring the real estate market and tax assessment for traded properties. The records are also used in private sectors, such as for the valuation of collaterals for loan approval.

The dataset comprises the lots sold in Gwacheon-si from 2016 to 2018, and includes the attributes of 1,111 samples. These attributes include sales price, zone improvement plan (ZIP) code, lot shape, and bearing. The following variables are used for the clustering analysis: sales price, longitude and latitude (hereafter denoted as X-coordinate and Y-coordinate, respectively), ZIP code, and zone. By employing these five variables, the lots close to each other in terms of sales price,

geographical location, and zoning are expected to group together to form a submarket. Although more variables, such as lot shape (regular or irregular) and bearing, are available as inputs, they reflect the physical characteristics of individual lots, rather than neighborhood/submarket characteristics; thus, these other variables are not employed for market delineation. Table 1 presents the descriptive statistics of sales samples.

Table 1 Descriptive statistics of the 1,111 lots sold from 2016 to 2018

| Variable | Min. | Mean | Median | Max. |
|---|---|---|---|---|
| Sales price (KRW/$m^2$) | 6,938 | 1,539,433 | 379,753 | 22,783,807 |
| ZIP code (25 levels) | #13820: 245 (22.1%)<br>#13840: 164 (14.8%)<br>#13814: 101 (9.1%)<br>#13813: 81 (7.3%)<br>#13801: 75 (6.8%)<br>#13824: 73 (6.6%) | | | |
| Zone (7 levels) | Green Belt: 951 (85.6%)<br>Residence 3: 103 (9.3%)<br>Residence 5: 40 (3.6%)<br>Natural & Green area: 8 (0.7%)<br>Residence 2: 6 (0.5%)<br>Residence 4: 2 (0.2%)<br>Residence 1: 1 (0.1%) | | | |

*Note: Only primary levels are presented in the ZIP code for readability.

The median sales price of lots in Gwacheon-si is 379,753 KRW/$m^2$, and most area is zoned for the Green Belt (85.6%). As shown in the table, ZIP code and zone are categorical variables. Particularly, the ZIP code is highly cardinal with 25 levels. All continuous variables (sales price, X-coordinate, and Y-coordinate) are scaled to have a mean of zero and a standard deviation of one before being fed into the clustering algorithm. A portion of the dataset (222 samples, 20%) is reserved to evaluate the performance of clustering algorithms.

The land sales records are collected by local governments and released to the public on a monthly basis. The records can be utilized to detect overheating spots in real estate submarkets and diagnose the sustainability of the overall market. In the future, local governments need to collect and agglomerate the land sales data in a more frequent cycle, for example, on a weekly or daily basis, to adapt to a rapidly changing real estate market. In addition to optimizing the data accumulation process, data-driven algorithms based on machine learning must be developed and deployed in administration tasks ranging from disclosing the real estate market in a transparent manner to detecting fraudulent land transactions.

### *3.2. Creating embedding vectors*

In the case of a high-cardinality categorical variable, such as the ZIP code, a one-hot encoding technique cannot be efficient in handling a large number of elements in a variable; thus, the high-cardinality categorical variable should be represented in the form of embedding vectors. The embedding vectors utilized in this study are obtained from the training results of a neural network.

The aforementioned network is a fully connected layer network with the following architecture: four input variables are employed, and then two embedding layers corresponding to the categorical variables are additionally specified and added to the network. An appropriate number of dimensions has to be determined for each embedding layer, and the prediction performance for various dimension sizes is reviewed through the usual cross-validation process. The number of dimensions assigned to each categorical variable through this cross-validation process is 10 and 3 for the ZIP code and zone, respectively. The number of dimensions in Fig. 2 is selected based on the results of the heuristic grid search. The grid search uses five-fold cross-validation to evaluate the possible combinations of values (the number of dimensions for the ZIP code and zone in this study), and the mean squared error (MSE) between the observed land prices and the predicted prices is used as a criterion for

the evaluation. As a rule of thumb, half the number of original levels is often used as a reference number for dimensions [39]. The ZIP code comprises 25 levels, and the zone consists of 7 levels, as shown in Fig. 2. Thus, according to this rule of thumb, good candidate numbers of dimensions for the ZIP code and zone are 12-13 and 3-4, respectively. Hence, the number of dimensions for the ZIP code is chosen as 10, and that for the zone is chosen as three, considering the results of the heuristic grid search and the rule of thumb together. Finally, to include more parameters to capture minor data nuances, three hidden layers are added to the end of the architecture. The output layer with one neuron corresponds to the land price estimated by the neural network. The final architecture of the neural network used to obtain the embedding vectors is presented in Fig. 2.

The resultant embedding vectors assume the following form: a 25 × 10 matrix for the ZIP code and a 7 × 3 matrix for the zone. Table 2 presents the embedding vectors of the zone. Interpreting the learned embedding vectors always involves subjective judgments, and this study does not attempt to interpret their meanings because the primary goal is to reuse them in a subsequent clustering algorithm and achieve a performance better than those of the baseline algorithms.

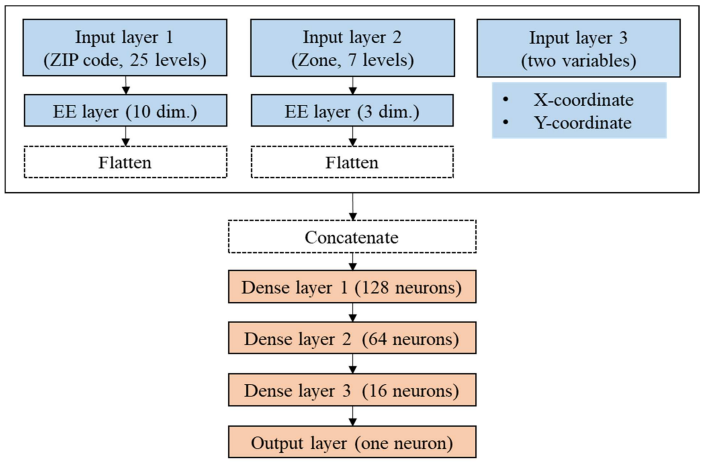



Fig. 2 Architecture of the neural network used to obtain embedding vectors

Table 2 Embedding vectors (7 × 3 matrix) of the zone learned from neural network training

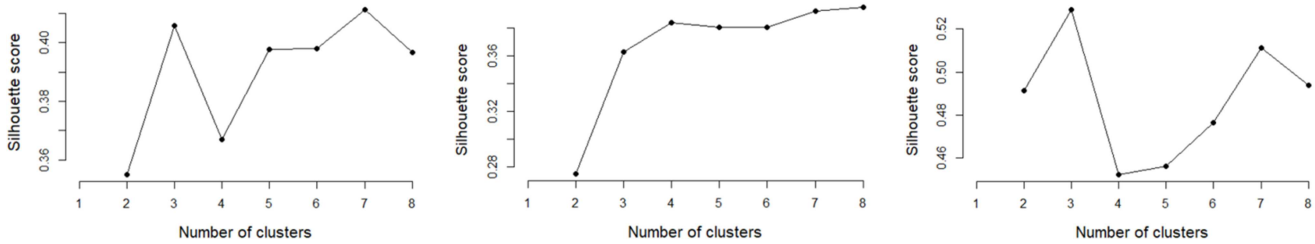
| Zone | Vector 1 | Vector 2 | Vector 3 |
|---|---|---|---|
| Residence 1 | 0.21375 | 0.20492 | 0.24875 |
| Residence 2 | -0.13061 | -0.18887 | -0.21148 |
| Residence 3 | 0.01262 | 0.01311 | 0.01660 |
| Residence 4 | 0.04791 | 0.04156 | -0.00550 |
| Residence 5 | -0.02817 | -0.02647 | -0.02822 |
| Natural & Green area | 0.00644 | -0.03270 | 0.01594 |
| Green Belt | 0.01215 | -0.02023 | -0.00067 |

### *3.3. Silhouette score*

When fitting a clustering algorithm such as *k*-means to a dataset, it is always subject to the judgment of a researcher to determine the optimal number of clusters, making the results vulnerable to criticism of subjectivity. Two popular methods are used to overcome this criticism: inertia and silhouette score analysis [40-41]. The former is defined as the mean squared distance between each observation and its closest centroid. The lower the inertia is, the better the algorithm is. However, this approach suffers from a limitation: as the number of clusters increases, the inertia always becomes lower. The silhouette score is a better measure to determine the number of clusters. It measures the extent of closeness between each observation in a cluster and the observations in the neighboring clusters, providing a way to assess the number of clusters. It varies between -1 and 1. A score close to 1 indicates that the observation is far from the neighboring clusters. A score close to 0 denotes that the observation is very close to the decision boundary between two neighboring clusters, and a negative score indicates that those observations might have been assigned to a wrong cluster. This study uses the silhouette score for the analysis.

A *k*-means clustering algorithm is used to delineate the real estate market, and three distance metrics are utilized. First, the Euclidean distance is applied for the five input variables, and categorical variables are converted to binary variables using the one-hot encoding approach. Second, the Gower distance, which is capable of dealing with both continuous and categorical variables, is used. Third, the Euclidean distance is applied again, although this time the categorical variables are converted to the continuous ones using entity embedding vectors. The continuous input variables (X- and Y- coordinates) are standardized before calculating the distance metrics. The *k*-means algorithm is implemented primarily following the clustering algorithm described in the work of Kaufman et al. [41].

Fig. 3 shows the respective silhouette score for *k*-means clustering based on the following three distance metrics: the Euclidean distance, the Gower distance, and the Euclidean distance empowered by embedding vectors (hereafter denoted as Euclidean distance 1, Gower distance, and Euclidean distance 2, respectively). For the categorical variables, the one-hot encoding approach is used in Euclidean distance 1, and the entity embedding approach is employed in Euclidean distance 2. After reviewing the visual depictions in the figure, three, four, and three clusters are chosen for each algorithm. As shown in panels (a) and (b) of Fig. 3, there are points with higher scores than those of the chosen points; however, the model sparsity is also considered when choosing the optimal number of clusters.



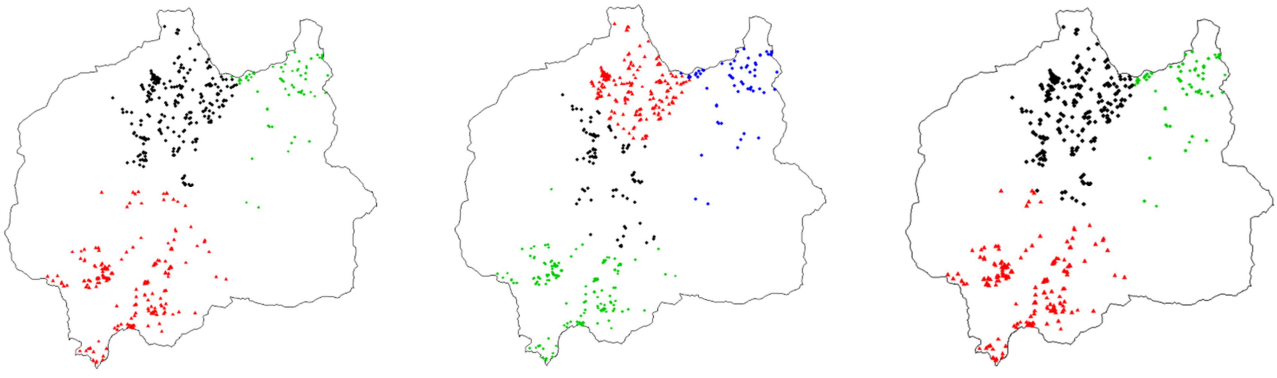

(a) Euclidean distance 1 (b) Gower distance (c) Euclidean distance 2

Fig. 3 Silhouette score for each clustering algorithm

## 4. Results

### *4.1. Results and evaluation*

Fig. 4 shows the market delineation results from the three *k*-means algorithms based on the Euclidean distance 1, Gower distance, and Euclidean distance 2, respectively. Each *k*-means algorithm produces three, four, and three submarkets, respectively, as shown in the panels (a), (b), and (c) in the figure. The three submarkets delineated by the *k*-means algorithm based on Euclidean distances 1 and 2 appear to correspond approximately to northern, southern, and eastern parts of the study area. The four submarkets identified by the *k*-means algorithm utilizing Gower distance correspond approximately to northern, middle, southern, and eastern parts of the study area.



(a) Euclidean distance 1 (b) Gower distance (c) Euclidean distance 2

Fig. 4 Market delineation results

The clustering results generally need to be checked using external validity measures, which can vary depending on the dataset and application areas, as explained earlier. This study uses the predictive accuracy of price estimation for the external validity measure. Conversely, the delineation results are evaluated on the basis of the accuracy of the predicted price for each submarket. This approach is often used to compare the resultant submarkets delineated by clustering algorithms [42]. The price is estimated using the well-established hedonic pricing model, as follows [43-45]:

$$Price_i = ZIP_i + Zone_i + Year_i + Shape_i + Bearing_i + Area_i \tag{3}$$

where $Price_i$ denotes the sales price per square meter of lot $i$, $ZIP_i$ and $Zone_i$ denote the areas to which lot $i$ belongs, and $Year_i$ denotes the year in which lot $i$ is sold. $Shape_i$ of lot $i$ has two levels: regular and irregular shape. $Bearing_i$ of lot $i$ comprises four levels: east, west, south, and north. Finally, $Area_i$ represents the size of lot $i$ measured in square meters.

Then, the root-mean-square error (RMSE) criterion is used for comparing the predictive accuracy, as follows [46]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2} \tag{4}$$

where $\hat{y}$ denotes the price predicted by the pricing model, and $y$ denotes the observed price.

Table 3 presents the predictive accuracy (RMSE) for each submarket. Prices in which RMSE is measured are standardized to have a mean of zero and a standard deviation of one. Although the marginal difference in average RMSE exists between Euclidean distance 1 and Gower distance, the average RMSE for the submarkets created by Euclidean distance 2 is significantly reduced compared with the former two results. The average RMSE in Euclidean distance 2 is reduced by 10% approximately, from 0.076-0.077 (Euclidean distance 1 and Gower distance) to 0.069. This decrease could be attributed to the capability of embedding vectors used in Euclidean distance 2 to extract the intrinsic relationships between levels in a categorical variable. Conversely, by identifying the meaningful patterns inherent in categorical data and representing the patterns in the form of numerical vectors, the entity embedding approach can enhance the relevance of delineated submarkets.

Table 3 Predictive accuracy for each submarket constructed by Euclidean distance 1, Gower distance, and Euclidean distance 2, respectively

| RMSE | Euclidean distance 1 | Gower distance | Euclidean distance 2 |
|---|---|---|---|
| Submarket 1 | 0.075 | 0.074 | 0.071 |
| Submarket 2 | 0.104 | 0.088 | 0.090 |
| Submarket 3 | 0.049 | 0.097 | 0.047 |
| Submarket 4 | - | 0.048 | - |
| **Average** | **0.076** | **0.077** | **0.069** |

### *4.2. Interpreting the resultant submarkets*

The average RMSE for the submarkets constructed by Euclidean distance 2 is the lowest. The submarkets are illustrated in Fig. 5, which shows the subway line and arterial road. The clustering algorithm agglomerated the individual sales lots in Gwacheon-si into three submarkets: northern, southern, and eastern submarkets. The northern submarket, which has old single-family houses, is a typical residential area with a well-established urban infrastructure. The southern submarket can be characterized by the landscapes mixed with public facilities and apartments; the public facilities include Gwacheon City Hall and the old Central Government Complex. The eastern submarket mainly comprises low hills, small mountains, and sparsely located houses.

In Fig. 5, the area indicated by the dotted circle is classified as belonging to the southern submarket by Euclidean distance 1, but is reclassified as belonging to the northern submarket when Euclidean distance 2 is applied. It is the area located close to the subway line and appears to be difficult to delineate in a confident manner. Consultations from the local experts, such as real

estate brokers and property appraisers, also confirm the ambiguity of the submarket membership of this area. Some experts classify the dotted area as belonging to the northern submarket in Gwacheon-si, while others consider that the area should belong to the southern submarket. Although the topic of the submarket membership of the area is controversial even among domain experts, the data-driven clustering by Euclidean distance 2 labels the area as belonging to the northern submarket in Gwacheon-si, and the validity is proved by the lowest RMSE obtained from a hedonic pricing model.

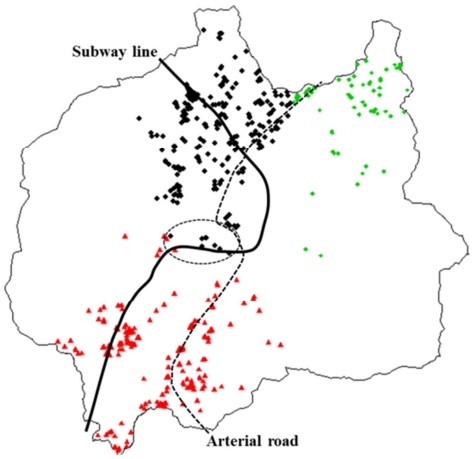



Fig. 5 Submarkets constructed by Euclidean distance 2

## 5. Conclusions

A clustering algorithm becomes inefficient or unstable when categorical variables are dominant in the input dataset, and the situation worsens in the case of high-cardinality categorical variables. This study attempted to enhance the performance of a clustering algorithm by employing an entity embedding approach to high-cardinality variables.

Gwacheon-si was chosen for the analysis, and a clustering algorithm was applied to the sales records of lots to delineate the real estate market. Embedding vectors for the ZIP code and zone were learned from neural network training and subsequently applied to the clustering algorithm. The results showed that the submarket delineation created by the Euclidean distance-based clustering algorithm equipped with embedding vectors outperformed the ones achieved by baseline models, such as the ordinary Euclidean distance-based algorithm and the Gower distance-based algorithm.

This study offered an efficient alternative to handle these categorical variables when applying clustering algorithms: the clustering analysis results can be improved by using embedding vectors learned from neural network training, which has been utilized universally in machine learning applications that handle unstructured data. This study might promote the rapid adoption of machine-learning tools in the field of structured data.

The dataset used in the study comprised 1,111 lots, but the data size may be insufficient to draw a generalizable conclusion. Different results would possibly be observed if the proposed approach was applied to different datasets or different local areas. Empirical experiments at a more extensive scale need to be attempted in future studies.

## Conflicts of Interest

The author declares no conflict of interest.

## References

[1] V. Goyal, G. Singh, O. Tiwari, S. Punia, and M. Kumar, “Intelligent Skin Cancer Detection Mobile Application Using Convolution Neural Network,” Journal of Advanced Research in Dynamical and Control Systems, vol. 11, no. 7, pp. 253-259, 2019.

[2] A. Aggarwal, M. Alshehri, M. Kumar, P. Sharma, O. Alfarraj, and V. Deep, "Principal Component Analysis, Hidden Markov Model, and Artificial Neural Network Inspired Techniques to Recognize Faces," Concurrency and Computation: Practice and Experience, vol. 33, no. 9, e6157, May 2021.
[3] M. Alshehri, M. Kumar, A. Bhardwaj, S. Mishra, and J. Gyani, "Deep Learning Based Approach to Classify Saline Particles in Sea Water," Water, vol. 13, no. 9, 1251, 2021.
[4] A. Aggarwal, A. Rani, P. Sharma, M. Kumar, A. Shankar, and M. Alazab, "Prediction of Landsliding Using Univariate Forecasting Models," Internet Technology Letters, in press.
[5] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Cambridge: Morgan Kaufmann, 2016.
[6] A. C. Goodman and T. G. Thibodeau, "Housing Market Segmentation and Hedonic Prediction Accuracy," Journal of Housing Economics, vol. 12, no. 3, pp. 181-201, September 2003.
[7] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties," Biometrics, vol. 27, no. 4, pp. 857-871, December 1971.
[8] L. R. Dice, "Measures of the Amount of Ecologic Association between Species," Ecology, vol. 26, no. 3, pp. 297-302, July 1945.
[9] P. Legendre and L. Legendre, Numerical Ecology, Burlington: Elsevier Science, 2012.
[10] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283-304, 1998.
[11] S. S. Khan and A. Ahmad, "Cluster Center Initialization Algorithm for K-Modes Clustering," Expert Systems with Applications, vol. 40, no. 18, pp. 7444-7456, December 2013.
[12] N. Sharma and N. Gaud, "K-Modes Clustering Algorithm for Categorical Data," International Journal of Computer Applications, vol. 127, no. 17, pp. 1-6, October 2015.
[13] C. Guo and F. Berkhahn, "Entity Embeddings of Categorical Variables," https://arxiv.org/pdf/1604.06737.pdf, April 22, 2016.
[14] V. Efthymiou, O. Hassanzadeh, M. Rodriguez-Muro, and V. Christophides, "Matching Web Tables with Knowledge Base Entities: From Entity Lookups to Entity Embeddings," International Semantic Web Conference, pp. 260-277, October 2017.
[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation," Conference on Empirical Methods in Natural Language Processing, pp. 1532-1543, October 2014.
[16] O. Abdelwahab and A. Elmaghraby, "UofL at SemEval-2016 Task 4: Multi Domain Word2vec for Twitter Sentiment Classification," 10th International Workshop on Semantic Evaluation, pp. 164-170, June 2016.
[17] Z. Chen, Y. Huang, Y. Liang, Y. Wang, X. Fu, and K. Fu, "RGloVe: An Improved Approach of Global Vectors for Distributional Entity Relation Representation," Algorithms, vol. 10, no. 2, 42, 2017.
[18] M. Aydoğan and A. Karci, "Turkish Text Classification with Machine Learning and Transfer Learning," International Artificial Intelligence and Data Processing Symp., pp. 1-6, September 2019.
[19] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," International Conference on Machine Learning, pp. 478-487, June 2016.
[20] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved Deep Embedded Clustering with Local Structure Preservation," 26th International Joint Conference on Artificial Intelligence, pp. 1753-1759, August 2017.
[21] C. Wu and R. Sharma, "Housing Submarket Classification: The Role of Spatial Contiguity," Applied Geography, vol. 32, no. 2, pp. 746-756, March 2012.
[22] B. Keskin and C. Watkins, "Defining Spatial Housing Submarkets: Exploring the Case for Expert Delineated Boundaries," Urban Studies, vol. 54, no. 6, pp. 1446-1462, 2017.
[23] S. Openshaw, "A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling," Transactions of the Institute of British Geographers, vol. 2, no. 4, pp. 459-472, 1977.
[24] D. P. Claessens, S. Boonstra, and H. Hofmeyer, "Spatial Zoning for Better Structural Topology Design and Performance," Advanced Engineering Informatics, vol. 46, 101162, October 2020.
[25] R. M. Assunção, M. C. Neves, G. Câmara, and C. da Costa Freitas, "Efficient Regionalization Techniques for Socio-Economic Geographical Units Using Minimum Spanning Trees," International Journal of Geographical Information Science, vol. 20, no. 7, pp. 797-811, 2006.
[26] W. Lin and Y. Li, "Parallel Regional Segmentation Method of High-Resolution Remote Sensing Image Based on Minimum Spanning Tree," Remote Sensing, vol. 12, no. 5, 783, 2020.

[27] Z. Cai, J. Wang, and K. He, "Adaptive Density-Based Spatial Clustering for Massive Data Analysis," IEEE Access, vol. 8, pp. 23346-23358, 2020.
[28] N. Jabeur, A. U. H. Yasar, E. Shakshuki, and H. Haddad, "Toward a Bio-Inspired Adaptive Spatial Clustering Approach for IoT Applications," Future Generation Computer Systems, vol. 107, pp. 736-744, June 2020.
[29] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," Journal of the American Statistical Association, vol. 66, no. 336, pp. 846-850, December 1971.
[30] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, November 1987.
[31] S. Eldridge, D. Ashby, C. Bennett, M. Wakelin, and G. Feder, "Internal and External Validity of Cluster Randomised Trials: Systematic Review of Recent Trials," British Medical Journal, vol. 336, 876, April 2008.
[32] M. Rezaei and P. Fränti, "Set Matching Measures for External Cluster Validity," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 8, pp. 2173-2186, August 2016.
[33] X. Li, W. Liang, X. Zhang, S. Qing, and P. C. Chang, "A Cluster Validity Evaluation Method for Dynamically Determining the Near-Optimal Number of Clusters," Soft Computing, vol. 24, no. 12, pp. 9227-9241, 2020.
[34] S. S. Kumar, S. T. Ahmed, P. Vigneshwaran, H. Sandeep, and H. M. Singh, "Two Phase Cluster Validation Approach Towards Measuring Cluster Quality in Unstructured and Structured Numerical Datasets," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 7, pp. 7581-7594, 2021.
[35] C. A. Lipscomb and M. C. Farmer, "Household Diversity and Market Segmentation within a Single Neighborhood," The Annals of Regional Science, vol. 39, no. 4, pp. 791-810, December 2005.
[36] Y. Tu, H. Sun, and S. M. Yu, "Spatial Autocorrelations and Urban Housing Market Segmentation," The Journal of Real Estate Finance and Economics, vol. 34, no. 3, pp. 385-406, 2007.
[37] Z. Liu, J. Cao, R. Xie, J. Yang, and Q. Wang, "Modeling Submarket Effect for Real Estate Hedonic Valuation: A Probabilistic Approach," IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 7, pp. 2943-2955, July 2021.
[38] KOSTAT, "Statistics Korea: Population and Households," http://kostat.go.kr/portal/eng/pressReleases/8/1/index.board, 2020.
[39] A. Koul, S. Ganju, and M. Kasam, Practical Deep Learning for Cloud, Mobile, and Edge: Real-World AI and Computer-Vision Projects Using Python, Keras, and TensorFlow, Sebastopol: O'Reilly Media, 2019.
[40] A. Struyf, M. Hubert, and P. Rousseeuw, "Clustering in an Object-Oriented Environment," Journal of Statistical Software, vol. 1, no. 4, pp. 1-30, February 1997.
[41] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Hoboken: John Wiley & Sons, 2009.
[42] S. C. Bourassa, F. Hamelink, M. Hoesli, and B. D. MacGregor, "Defining Housing Submarkets," Journal of Housing Economics, vol. 8, no. 2, pp. 160-183, June 1999.
[43] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," Journal of Political Economy, vol. 82, no. 1, pp. 34-55, January-February, 1974.
[44] S. Catma, "The Price of Coastal Erosion and Flood Risk: A Hedonic Pricing Approach," Oceans, vol. 2, no. 1, pp. 149-161, March 2021.
[45] P. M. Campos, J. S. Thompson, and J. P. Molina, "Effect of Irrigation Water Availability on the Value of Agricultural Land in Guanacaste, Costa Rica: A Hedonic Pricing Approach," e-Agronegocios, vol. 7, no. 1, pp. 38-55, 2020.
[46] D. Wackerly, W. Mendenhall, and R. L. Scheaffer, Mathematical Statistics with Applications, Belmont: Cengage Learning, 2014.