BUSINESS & DIPLOMACY REVIEW

1. évfolyam, 2. szám
2023/2

TARTALOM

Business & Diplomacy Review
Kiadó: Budapesti Gazdasági Egyetem

Szerkesztőség
1165 Budapest, Diósy Lajos u. 22-24.
https://uni-bge.hu/hu/business-diplomacy-review

Főszerkesztő
Sáringer János

Felelős szerkesztő
Csekő Katalin

Tipográfia és tördelés
Nagy és Hetyei Kft.

ISSN 3004-0116
HU ISSN 3004-0116

TARTALOM

Business & Diplomacy Review I. évfolyam | 2. szám | 2023. december

HIERARCHICAL CLUSTERING COMBINED WITH NEURAL NETWORKS

Árpád Bánhalmi

Abstract

This study presents how hierarchical clustering combined with neural networks operates and sheds light on how these models can be integrated with fundamental concepts. The hybrid model created can be applied to various tasks depending on the configuration of hyperparameters. The study serves a demonstration purpose by showing how the suggested model performs when run on a real dataset, where it attempts to estimate ordinal target variable values. The model's performance is analyzed in detail, with particular emphasis on overfitting, stability, and estimation accuracy. The hybrid model generates multiple estimates for the target variable values simultaneously - in this case, ten. These estimates are compared using available metrics, and then a method for deriving the final estimate is presented.

Keywords: neural network, hierarchical clustering, overfitting management

Introduction

In the context of machine learning models, a few hundred, a few thousand, or even tens of thousands of samples are considered a small sample size. Small-sample machine learning models encompass the procedures and methods for efficiently analyzing such limited data or rare events [1].

The establishment of small-sample machine learning models presents several challenges. There is a limited amount of data available, which makes it difficult to train and evaluate models. With small sample sizes, models are prone to overfitting, meaning they adapt too closely to the training dataset. Developing models capable of producing generalizable results from low-representative data due to data

scarcity is challenging. In the case of small sample sizes, noise and measurement errors can have a more significant impact on model performance. If certain classes are underrepresented due to the small sample size, models may tend to disregard them [2].

Combining hierarchical clustering with neural networks can offer new perspectives in the field of small-sample data analysis and machine learning, mitigating potential pitfalls [3]. Through this combination, it becomes possible to uncover the hierarchical structure of data and extract interpretable relationships that can be utilized with neural networks for further analysis and predictions [4].

Structuring Variables with Hierarchical Clustering

Hierarchical clustering is a method for grouping data in a step-by-step manner into clusters within a hierarchical structure. There are two main types of hierarchical clustering: agglomerative and divisive clustering. In agglomerative clustering, each data point initially belongs to its own cluster, and then the closest clusters are gradually merged until only one cluster remains. In contrast, divisive clustering starts with all data points in a single common cluster and then separates them. The result of both methods is a hierarchical tree structure that represents the data's grouping.

Hierarchical clustering allows for exploring both the hierarchy of individuals and the hierarchy of features (variables). In this case, we will be examining the hierarchical clustering of features (variables). In this scenario, the resulting dendrogram, which is a hierarchical tree, encodes the hierarchy of variables. The root of the dendrogram represents all features, where all variables are in a single large cluster. Internal nodes in the dendrogram depict the hierarchy of variables or variable groups based on their similarity. At the end of the dendrogram are the leaves at the lowest level of the hierarchical tree, which represent specific variables.

Various similarity measures can be applied in hierarchical clustering to determine the distance or similarity between clusters, such as Euclidean, Manhattan, cosine, Pearson correlation, Spearman rank correlation, Kendall tau rank correlation, Canberra distance, Chebyshev distance, Bray-Curtis distance, Hamming distance, Tanimoto

coefficient. These distance or similarity measures examine different characteristics and can be chosen based on the specific clustering task and the nature of the data. These measures are related to the linkage methods, such as complete linkage, single linkage, average linkage, or the Ward method, which determine how clusters are connected. The choice of linkage method depends on the data structure, the nature of variables, and their content [5].

Key Concepts Related to the Required Neural Network

Machine learning procedures can be categorized based on various aspects, such as learning mode, architecture, application domain, and operational principle. For the model to be presented, learning mode and architecture are particularly crucial, and these key concepts will be briefly explained. In machine learning, various architectures can be found, but the focus will be primarily on the difference between feedforward and deep feedforward neural networks. Feedforward neural networks are artificial neural networks where neurons are organized into layers, and data flow in one direction, forward [6]. These architectures are particularly well-suited for classification and regression tasks, as well as problems where the sequential order of input data is less significant.

Deep feedforward neural networks are a special type of feedforward network where the connections between layers can branch in multiple directions [7]. As a result, they enable parallel data processing in different branches and create higher-level hierarchical representations in hidden layers. Branching networks allow the handling of various aspects that later combine to generate output predictions. Such networks provide flexibility and efficiency because individual branches can be trained separately, and output layers can work independently, making them less sensitive to noise or missing data.

The depth of the architecture refers to the number of layers in the network. Deep neural networks are machine learning models that contain multiple hidden layers. These are ideal for solving complex problems that require higher-level features and hidden relationships in the input data. Deep networks result in a deeper hierarchy, enabling the exploration of more complex patterns [8].

The model to be presented is a deep feedforward branching neural network with deep hidden layers, and the architecture is supported by advanced regularization and normalization techniques.

Construction of a Neural Network Based on Hierarchical Clusters

Let us turn our attention to the interpretability aspects of neural networks built on hierarchical clustering. Interpretable models are often simpler and more easily understandable than complex, black-box models [9]. In the proposed model, the goal is to reduce the number of variables by using the hierarchical grouping of variables while preserving essential information. The expected outcome of clustering is to highlight significant variables, which can reduce the noise level during neural network training, thus enhancing the model's performance. Furthermore, the structured arrangement of variables can contribute to improving the model's performance and reducing the risk of overfitting, enhancing the model's generalization ability.

The construction of the neural network built on hierarchical clustering proceeds as follows:

In the first step, hierarchical clustering is performed based on the input variables. As a result, a hierarchical tree structure is created, which includes the grouping of variables at various levels. Based on the hierarchical clustering tree structure, the hidden layers of the neural network are positioned – the branches of the tree determine the relationships between the hidden layers. Then, a separate output layer is assigned to each hidden layer. During neural network training, the values of the input variables propagate forward through the branches created based on the hierarchical clustering tree structure. The neural network is trained using the standard error backpropagation method. The individual output layers follow the structure of the clusters, making it easy to identify which output layer is associated with which cluster and what predictions it generates.

Demonstration of a Neural Network Based on Hierarchical Clusters

During the research, the operation of the proposed model is demonstrated using real data. Data from observations of a group of university students are utilized, representing an entire class (approximately 600 individuals) and containing information related to the performance of a specific course. The study does not delve into the students' performance but is limited to the examination of the model's characteristics using the data.

To facilitate understanding of the model's operation, let's first examine the content of the variables. Based on the course requirements, students submitted 11 mandatory assignments electronically throughout the semester. Students were allowed to submit the same assignment multiple times, and only the best result was counted. Course grades are as follows: 0 - no signature, for students without a signature; for students with a signature, 1 - fail, 2 - pass, 3 - average, 4 - good, and 5 - excellent.

With the proposed model, we predict students' end-of-semester grades (output variable - y) based on the results of their assignments (input variables - x_0, x_1, x_2, ..., x_{10}). The input variables are measured on a ratio scale, while the target variable is measured on an ordinal scale. To ensure proper training and validation, the dataset is divided into three parts: training (70%), validation (15%), and testing (15%). The proposed model combines two methods and is constructed in two steps.

1. Clustering of Variables Using Hierarchical Clustering

Clustering is performed on the training data using a distance measure based on partial correlation coefficients between input variables. This is expected to reduce instability and inaccuracies arising from multicollinearity.

2. Merging Clusters Based on Maximum Distance Between Clusters ("Complete" Method) (Figure 1).

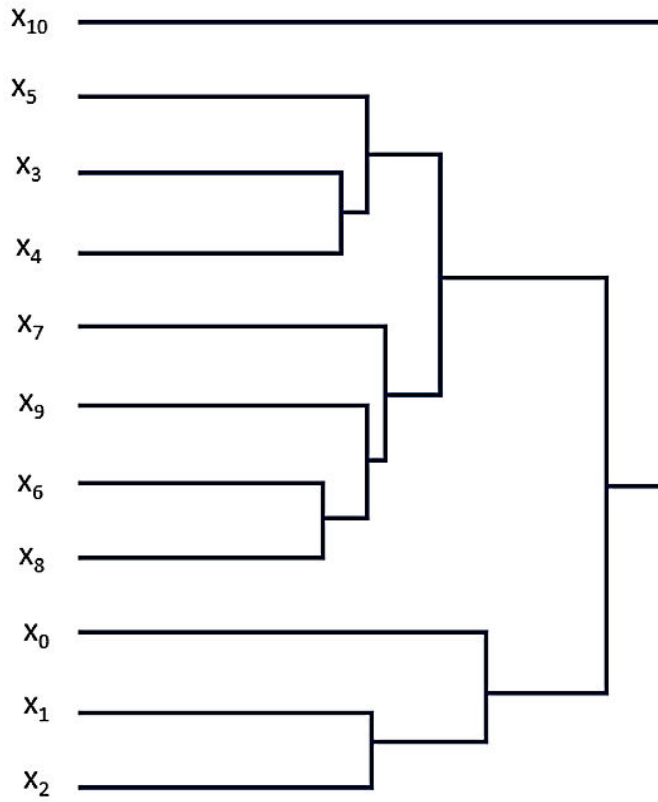

Figure 1: Illustration of Clusters of Input Variables with a Dendrogram. Source: Self-edited

Structure of the Neural Network Based on the Hierarchical Tree Structure

The neural network is constructed from the hierarchical tree structure with input variables x_0, x_1, ..., $x_{10.}$ The branching points of the tree determine the locations of the hidden layers, each of which has its corresponding output layer (Figure 2).

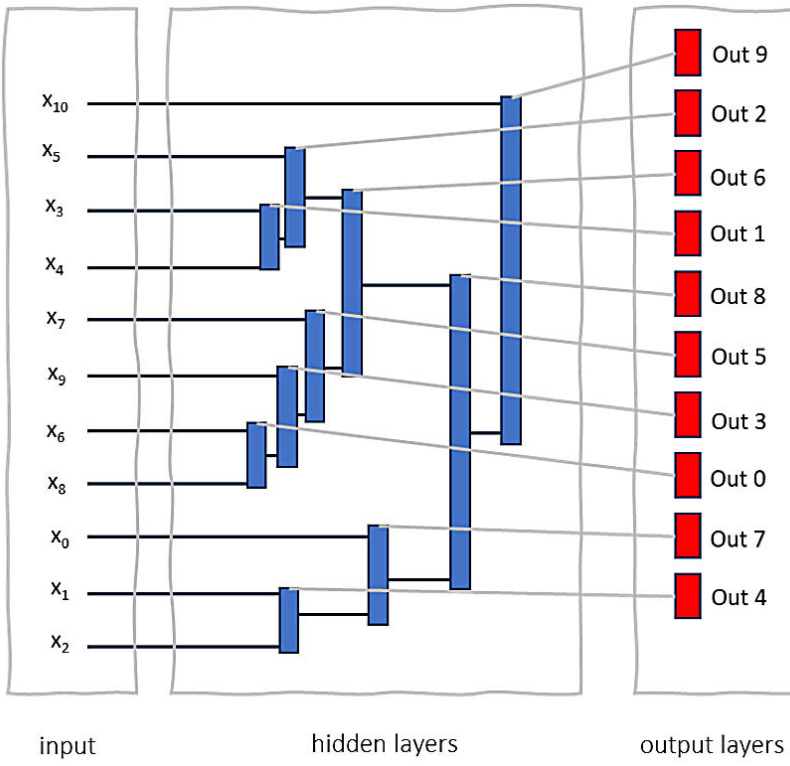

Figure 2: The Neural Network Constructed Based on Hierarchical Clustering. Source: Self-edited

The parameters of the neural network are as follows: The number of fully connected hidden layers is 10, with 50 neurons per layer and a softplus activation function. The number of fully connected output layers is 10, with 1 neuron per layer and an identity activation function, with output values rounded to the [0; 5] interval. Evaluation is performed with the Adam optimizer at a learning rate of 0.001, β_1 = 0.9, and β_2 = 0.999, using the Mean Squared Error (MSE) loss function. The number of epochs is set to 500.

The model's performance is measured using the ACC (accuracy), ACC' (modified accuracy), and CHI metrics. In the confusion matrix $\mathbf{C} = [c_{ij}]$, c_{ii} elements are on the main diagonal; $\mathbf{e}_i$ represents the i-th base vector, and denotes the submatrix obtained by removing the i-th row and j-th column of **C**. The distribution of the target variable values (0, 1, ..., 5) in the examined partitions (train, validation, test) is denoted by f_i (i, j = 0, 1, ..., 5).

$$n = \sum_{i=0}^{5} f_i$$

$$ACC = \frac{tr(C)}{n}$$

$$ACC' = \frac{tr\left(C(e_0^T C; Ce_5)\right) + tr(C) + tr\left(C(e_5^T C; Ce_0)\right)}{n}$$

$$CHI = \frac{\sum_{i=0}^{5} \frac{(c_{ii} - f_i)^2}{f_i}}{n}$$

The ACC is the well-known accuracy metric, indicating the proportion of elements correctly categorized according to the target variable. ACC' specifies the proportion of target variable values that were correctly estimated with an accuracy of ±1. CHI is the chi-squared statistic normalized by the number of data points, measuring the relationship between the main diagonal of the confusion matrix and the distribution of the target variable.

The loss values, which are the sum of the losses of the ten output layers, indicate a well-fitting model, with overfitting not occurring until the 500th epoch. The training and validation loss values reach a point of stability between the 100th and 200th epochs, and a healthy level of "generalization gap" can be observed (Figure 3). The cumulative loss values, which are the sum of the losses of the ten output layers, indicate a well-fitting model, with overfitting not occurring until the 500th epoch. The training and validation loss values reach a point of stability between the 100th and 200th epochs, and a healthy level of "generalization gap" can be observed (Figure 3).

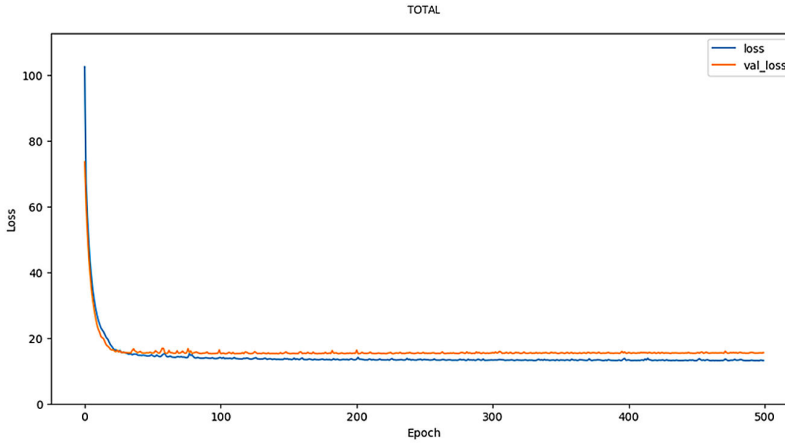

Figure 3: Total Learning Curve. Source: Self-edited

The input variables x_3, x_4, ..., x_9 belong to the same cluster based on the results of hierarchical clustering. In this cluster, the hidden layers correspond to the 0th, 1st, 2nd, 3rd, 5th, and 6th output layers (Figure 4).

Each learning curve of the output layers in this cluster resembles that of the 0th output layer (Figure 5). Based on this, it can be stated that all output layers in this cluster are free from overfitting and produce stable predictions. However, some of these predictions are limited in practical use. For instance, the 0th output layer does not provide predictions for the target variable with a value of 0, and the 1st output layer excludes the target variable's value of 5 from its predictions. The 2nd to 5th output layers exhibit behavior similar to the 0th and 1st output layers in terms of overfitting and generalization, but they do not provide predictions for one of the target variable's values.

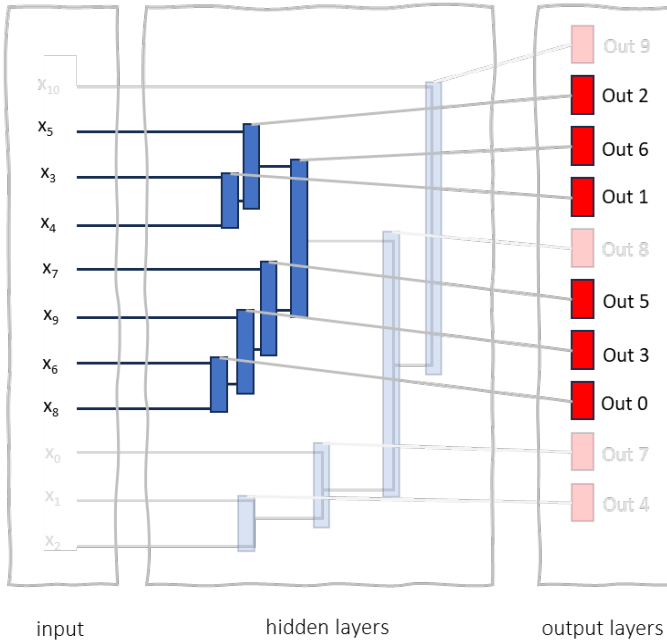

Figure 4: Highlighted Cluster of Variables x_3, x_4, ..., x_9 in the Neural Network. Source: Self-edited

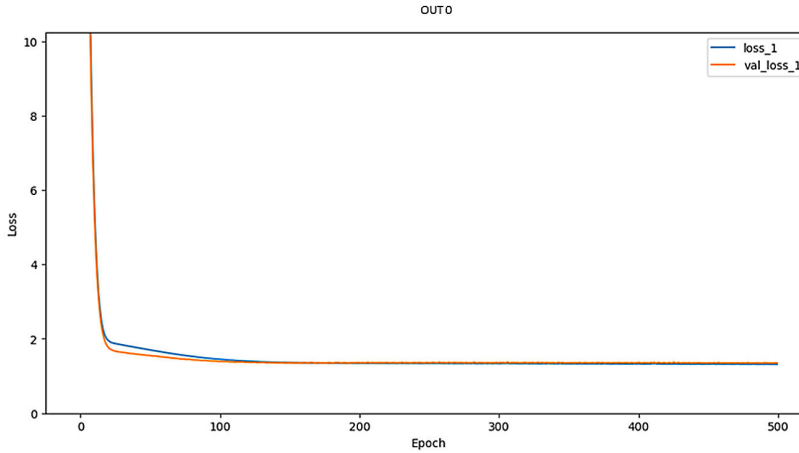

Figure 5: Learning Curve of Output Layer 0. Source: Self-edited

The 6th output layer behaves similarly in terms of stability, generalization, and overfitting as the previously discussed output layers (Figure 6). This output layer is connected to the deepest hidden layer in the examined cluster. However, this output layer provides predictions for every value of the target variable, though not necessarily with complete accuracy. Therefore, this output layer can be considered practical for making predictions. The accuracy of the predictions obtained from this layer can also be used because it predicts the target variable's values correctly in about 44% of cases (ACC = 0.4353) and within ±1 accuracy in 81% of cases (ACC' = 0.8118).

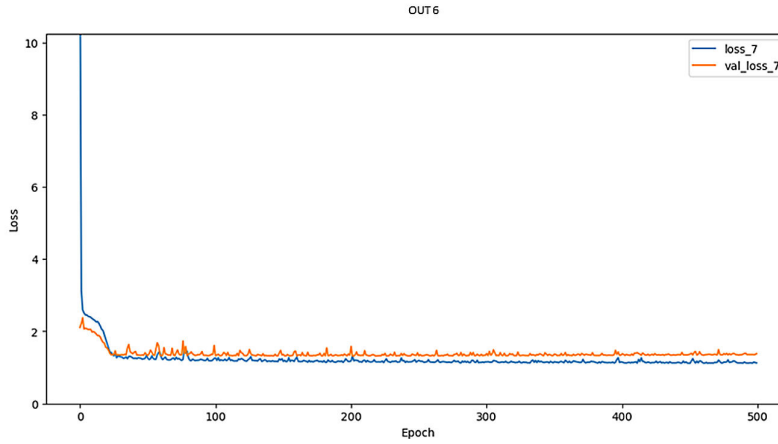

Figure 6: Learning Curve of Output Layer 6. Source: Self-edited

The evaluation of the CHI = 0.5890 metric in comparison to the same metric for the other output layers conveys information, as can be seen in Figure 7, the 6th output layer has the lowest value for this metric. Smaller CHI values indicate a better fit, so the predictions of the 6th output layer are the closest to perfect predictions among the examined output layers.

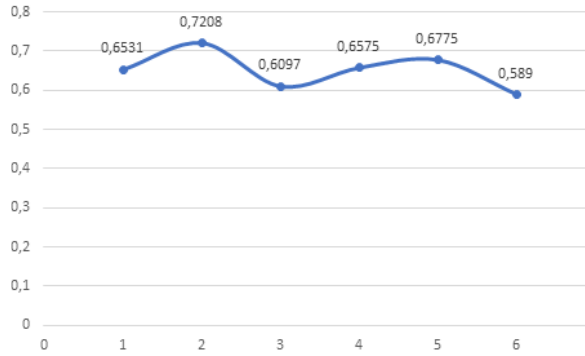

Figure 7: CHI Values for Individual Output Layers. Source: Self-edited

Examining the predictions of the output layers associated with these two additional clusters on the demonstration database reveals that we do not obtain more accurate or practically usable predictions than what we received in the case of the 6th output layer (Figures 8 and 9).

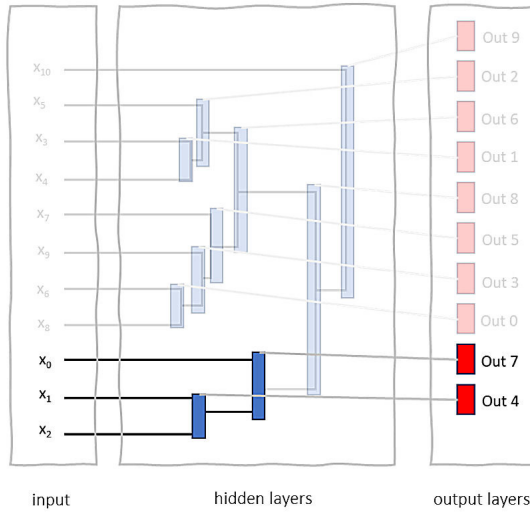

Figure 8: Second Cluster. Source: Self-edited

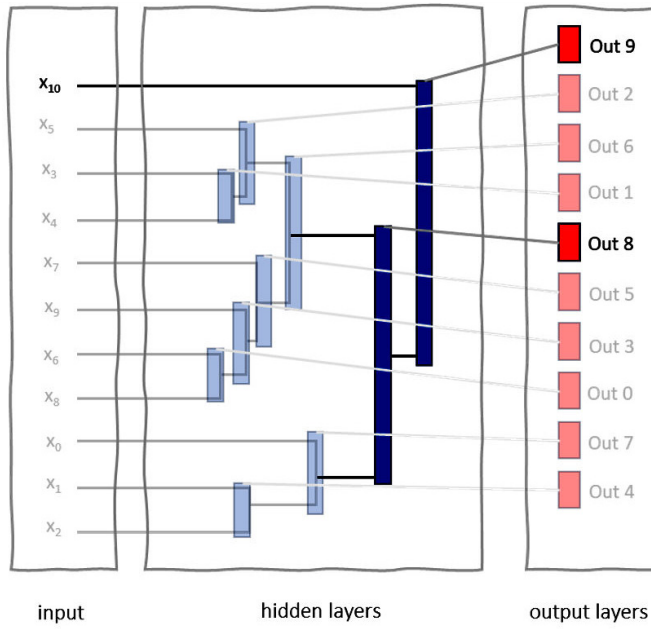

Figure 9: Third Cluster. Source: Self-edited

The developed model was successfully applicable to the presented dataset. Taking practical considerations into account, one could choose from the ten different estimations. Based on the metrics evaluating the estimations of the output layers (ACC, ACC', CHI), the output layer that could most efficiently and accurately estimate the target variable using this model was selected. The analysis of the learning curves showed that the ten parallel output layers were effectively capable of preventing overfitting, thus preserving the model's generalization ability.

Summary

The goal of combining hierarchical clustering and neural networks is to adapt our machine learning models to the complex structure of the data and hierarchical patterns. The branching architecture in the

hidden layers allows the model to interpret the data at different levels and discover complex relationships between them.

These hybrid models can be used for both regression and classification tasks, capable of predicting numerical values and determining classes or categories. These methods can also be successfully applied to ordinal target variables.

Parallel output layers are important in preventing overfitting. The separate output layers, based on the hierarchical tree structure, allow the model to operate sub-models on the cluster-specific inputs. This improves the model and sub-models' generalization ability and minimizes the risk of overfitting.

These models not only provide good performance but also assist in mapping and interpreting data structure. The hierarchical tree structure is particularly advantageous for tasks where data naturally organize hierarchically, such as biological systems, financial markets, or human resource management. Hierarchical clustering-based neural networks can be widely applied, including healthcare forecasting, financial portfolio management, consumer behavior analysis, environmental monitoring, and educational counseling.

In addition to the mentioned advantages, it's important to highlight that building and training such complex models is a significant challenge, requiring a substantial amount of high-quality data, expertise, and computational resources. With all these considerations, the hierarchical clustering combined with neural networks can be an outstanding tool for deeper data interpretation and the development of efficient predictive models.

References

1. Zhou, Shaoxuan (2023): An Analysis of The Small Sample Datasets Based on Machine Learning, Association for Computing Machinery, New York, NY, USA, ISBN 9781450397148, doi: 10.1145/3573428.3573720
2. Xu, P., Ji, X., Li, M. et al. (2023): Small data machine learning in materials science. npj Comput Mater 9, 42. https://doi.org/10.1038/s41524-023-01000-z

3. Zheng Y., Chen Q., Fan J., Gao X. (2020): Hierarchical convolutional neural network via hierarchical cluster validity based visual tree learning, Neurocomputing, Volume 409, Pages 408-419, ISSN 0925-2312
4. Zhuoyang L., Feng X. (2023): Interpretable neural networks: principles and applications, Frontiers in Artificial Intelligence, VOLUME 6, ISSN 2624-8212
5. Nielsen F. (2016): Introduction to HPC with MPI for Data Science, Springer Cham, Springer International Publishing Switzerland, ISBN 978-3-319-21902-8
6. Landau L. J., Taylor J. G. eds. (1998): Concepts for Neural Networks, A Survey, Springer-Verlag, London
7. López O. A. M., López A. M., Crossa J. (2022): Multivariate Statistical Machine Learning Methods for Genomic Prediction, Springer Cham, ISBN 978-3-030-89009-4
8. Taye, M.M. (2023): Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers 2023, 12, 91.
9. Hassija, V., Chamola, V., Mahapatra, A. et al. (2023): Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation (2023), https://doi.org/10.1007/s12559-023-10179-8

Business & Diplomacy Review I. évfolyam | 2. szám | 2023. december

Publikálás a Business & Diplomacy Review folyóiratban

A Folyóirat az alábbi témákhoz kapcsolódó kéziratokat fogad be lektorálásra:

- gazdaság,
- kereskedelem,
- pénzügyek,
- nemzetközi kapcsolatok,
- diplomácia,
- nemzetközi kommunikáció
- társadalomtudományok;
- közgazdasági tudományok.

A tanulmány absztraktja minden esetben angol nyelvű. A tanulmány magyar vagy angol nyelvű, amelynek terjedelme 30 000–40 000 leütés (az absztrakttal együtt).

A kéziratok előzetes befogadásának feltételei:

- a kézirat és annak szerzői megfelelnek a Folyóirat etikai szabályainak;
- a kézirat, illetve ahhoz tartalmában nagyon hasonló tanulmányt még nem publikálták;
- a benyújtott kézirat megfelel a formai követelményeknek.

A Folyóirat a szerzőknek tiszteletdíjat nem fizet. A Folyóirat minden egyes befogadott kézirat esetében kettős vaklektorálást alkalmaz, ami azt jelenti, hogy az anonimalizált anyagot a szerzők által nem ismert lektorok értékelik. A Folyóirat csak abban az esetben fogad be kéziratot publikálásra, ha azt mind a két vaklektor publikálásra ajánlja, és vaklektor(ok) által kért javításokat/kiegészítéseket a szerző(k) végrehajtotta/ák. Amennyiben az egyik lektor javításokkal publikálásra ajánlja a kéziratot, míg a másik nem, akkor a javítások után a témában jártas újabb vaklektornak kell értékelnie az anyagot. Akkor minősül egy tanulmány tartalmában nagyon hasonlónak egy korábbi tanulmányhoz képest, ha azok egyezősége 60% felett van. A szerzők minden egyes esetben kötelesek a vaklektorok által írt kifogásokra/javaslatokra tételesen írásban reagálni.

Kiadja: Budapesti Gazdasági Egyetem, Budapest

ISSN 3004-0116
HU ISSN 3004-0116

Kiadja: Budapesti Gazdasági Egyetem, Budapest

ISSN 3004-0116
HU ISSN 3004-0116