Embeddings for the Graph Classifier of Economic Activities

N. Blokhin, A. Vyrodov, S. Makrushin

Financial University under the Government of the Russian Federation, Moscow

SVMakrushin@fa.ru

A proper representation of the subject's industry sector in a key factor for solving a wide range of applied economic problems ranging from analysis of the enterprises' creditability to searching for the companies in the same market place. At this point All-Russian Classification of Economic Activities (OKVED) is used as the main industry sector description tool in Russia. [1]. This is a well-developed hierarchical system of economic activities. However, just as any hierarchical classifier, OKVED is limited to the small set of criteria used for building a hierarchy and determination of the proximity of hierarchy elements.

Modern methods of network analysis can be used to create a tool that gives more comprehensive determination of the proximity between industries. In particular, vector representations in a multidimensional space (embeddings) of network nodes might be applied for this purpose. The innovation proposed by the authors is an approach to building a network of economic activity codes and a method of building embeddings for nodes of the constructed network.

The first step to solve the problem of building embeddings is to create a network where the nodes represent the codes of economic activities from OKVED. The authors consider two types of links between the codes. The first type reflects the hierarchy of codes in OKVED. The second type complements the OKVED hierarchy with information about the proximity of economic activities based on ownership relationships between real companies. The Internet resource "Directory of organizations, enterprises and companies in Russia" [2] was used as an open source of information about companies, and the data about several tens of thousands of companies was obtained from it.

Recently, many methods of building embeddings for network nodes have been proposed. At this point methods of building embeddings and solving other graph problems based on deep learning models, namely graph neural networks (GNN), have become increasingly widespread [3]. Graph convolution neural networks (GCN) [4, 5, 6] which generalize the convolution operation on graphs are considered to be one of the most prominent directions of development in the field of graph neural networks. The authors use this specific approach for the embedding construction in the work. The problem of link prediction for the second type of relationships based on the created graph was resolved in order to obtain embeddings of economic activity codes.

These obtained embeddings contain information both about the relationships between the codes provided by OKVED and the proximity of codes, taking into account the links between real companies that are representatives of industries. Visualization of the trained embeddings is shown in Fig. 1. In Fig. 1 and Fig. 2, one can see that codes of similar economic activities are located close to each other in the embedding space despite the fact that some of them are not neighbors in OKVED.

The embeddings obtained in the work can be used in machine learning models both directly as features describing the company's industry sector, and for determination the industry proximity between enterprises.

References

1. All-Russian Classification of Economic Activities " (approved By order of Rosstandart dated от 31 January 2014 N 14) (editon 12 August 2021)
2. Directory of organizations, enterprises and companies in Russia - List-Org , accessed 23 November 2021, < https://www.list-org.com >
3. J. Zhou, G. Cui et al. Graph neural networks: A review of methods and applications. arXiv preprint arXiv:1812.08434, 2018.
4. T. Kipf, M. Welling. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR), 2017.
5. M. Schlichtkrull, T. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling. Modeling relational data with graph convolutional network. ESWC. Springer, 2018, pp. 593–607.
6. W. Hamilton, Z. Ying, J. Leskovec. Inductive representation learning on large graphs. In Proc. of NIPS, 2017, pp. 1024–1034.

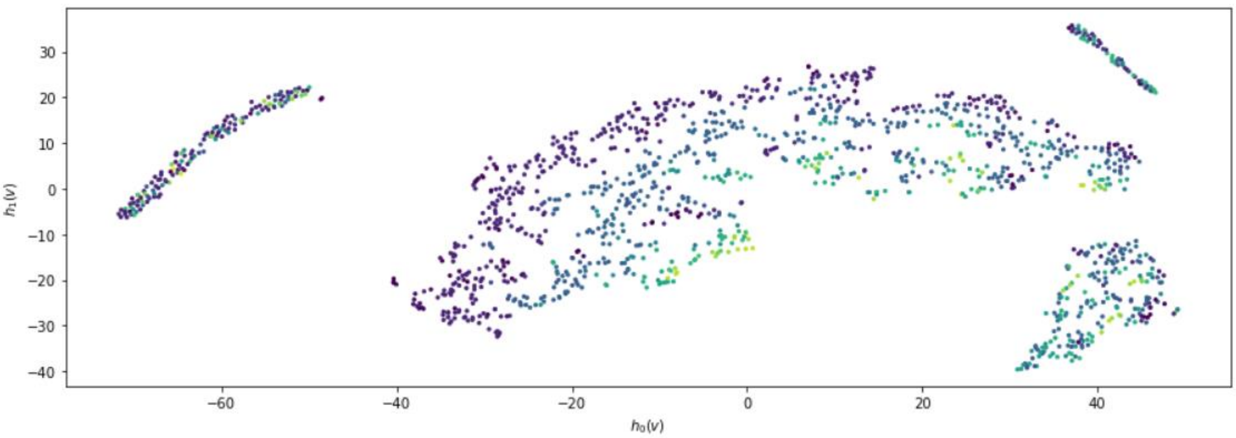

Fig. 1. Visualization of embeddings of codes of economic activity using the TSNE algorithm for dimensionality reduction. The color indicates the section of the economic activity code.

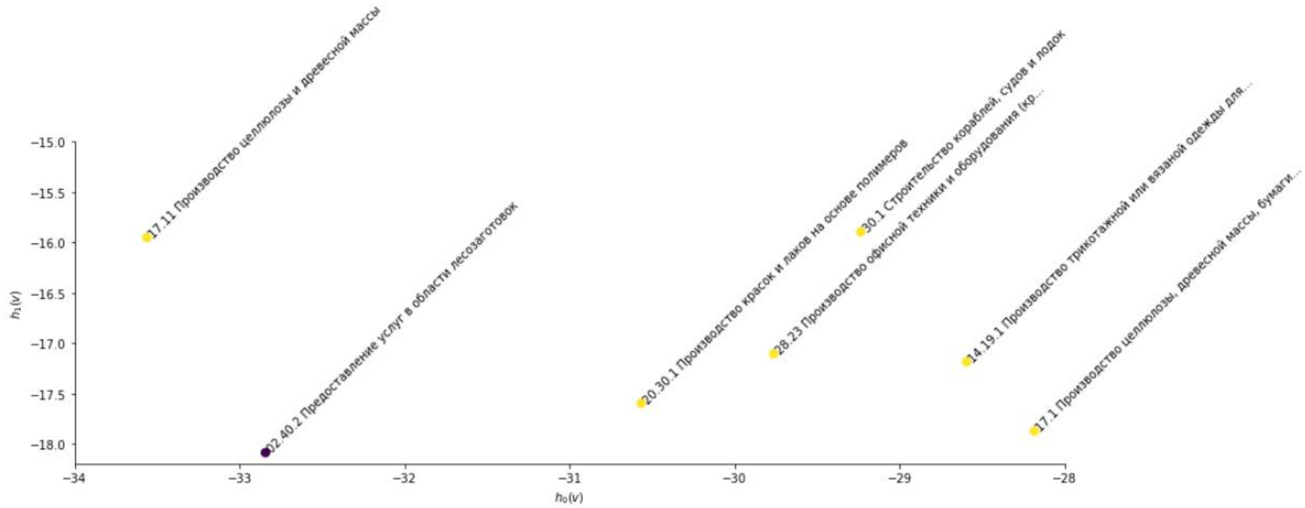

Fig. 2. An example of the location of codes related to the type of activity (related to wood processing and the production of wood goods) in the embedding space.