Federal state autonomous
educational institution for higher education
«National research university
«Higher school of economics»
Faculty of Graduate School of Business
Master's Programme
«Business Analytics and Big Data Systems»

Project

ON THE TOPIC

"LastFM Users Country Prediction"

Performed by 1 course students,
Shirnin Alexander, Markin Nikita, Sukharkov Alexander

Moscow 2023

Contents

Abstract

In this project, we aim to predict the country of Last.fm users based on the artists they like and their social connections. We plan to use machine learning and graph-based deep learning methods, including graph neural networks (GNNs), to learn meaningful representations of users and their relationships in the graph. By leveraging these representations, we aim to develop accurate prediction models for each user's country. We will compare the performance of these methods in terms of accuracy and efficiency to determine the most effective approach for predicting user countries in Last.fm. The most effective approach has the potential to improve user profiling and recommendation systems, which can enhance the user experience on the platform.

Key words - GNN, Gradient boosting, Embedding Space, Node2Vec.

1 Introduction

Music is an essential part of human life and culture, and it has become even more accessible in the digital age. Last.fm was one of the largest music streaming platforms in the world, with millions of users and vast amounts of data about their listening habits. Understanding the preferences and behaviors of Last.fm users can provide valuable insights into the global music scene and help improve personalized music recommendations.

In this research project, we aim to predict the country of Asian Last.fm users based on the users they follow and the musicians they like. We believe that music preferences can reveal important information about a person's cultural background, and that this information can be used to accurately predict a user's country of origin. By analyzing the listening habits of Asian users on Last.fm, we can gain a deeper understanding of the music tastes and cultural differences in the region.

We are going to use the LastFM Asia Social Network dataset [1]. It consists of 7,624 nodes and has 27,806 edges. Each node has a label, that represents a country. There are 18 countries in total.

We plan to use graph-based machine learning techniques to analyze the dataset and develop predictive models. Our approach will involve feature extraction from the graph structure, that is incorporating embeddings of the artist preferences of each user. We will evaluate the performance of our models using the accuracy metric and compare various methods to analyze them.

The outcomes of our research can have significant implications not only for the music industry but also beyond it. By developing accurate predictive models, our work can aid music streaming platforms in enhancing their recommendation systems, providing users with more personalized and relevant content. We will understand, which techniques are working better for this problem.

2 Trainnig

2.1 Anticipated Methods

We propose three different approaches to solve that problem:

- **Node2vec embeddings + Catboost, XGBoost, or LGBM**: We utilize the Node2vec algorithm to generate embeddings for the graph nodes. These embeddings capture the structural information of the graph and the relationships between users. We then employ Catboost, XGBoost, or LGBM as regression models on these embeddings to predict the country of Last.fm users.

- **SVD on graph adjacent matrix + Catboost, XGBoost, or LGBM**: We perform Singular Value Decomposition (SVD) on the graph's adjacency matrix to obtain embeddings for each node. These embeddings represent the latent features of the users within the graph. Another option is to concatenate this matrix with the feature matrix of users. We subsequently employ Catboost, XGBoost, or LGBM as regression models using these embeddings to predict the country labels.

- **Graph Attention Neural Network (GAT)**: We employ a Graph Attention Neural Network, a deep learning model specifically designed for graph-structured data. The GAT model captures the relational information within the graph and learns node embeddings that can be used for country prediction.

2.2 Dataset Split and Hyperparameter Tuning

To ensure an unbiased evaluation of the predictive models, we divided the LastFM Asia Social Network dataset into three subsets: 80% for training and validation, and the remaining 20% for testing. This division allows us to assess the performance of the models on unseen data and generalize their effectiveness. Furthermore, from the training dataset we utilized the 10% as a validation dataset to tune the hyperparameters of the models. This step ensures that the models are optimized for performance by selecting the most suitable hyperparameters. By following this rigorous dataset splitting and hyperparameter tuning approach, we can confidently conclude that the reported accuracies are reliable indicators of the models' predictive capabilities.

3 Results

Method	Accuracy (%)
Node2vec + LGBM	87.0
Node2vec + Catboost	87.2
Node2vec + XGBoost	86.6
SVD + LGBM	70.1
SVD + Catboost	61.6
SVD + XGBoost	78.3
GAT	**90.6**

Table 1: Accuracy results for LastFM Users Country Prediction

The results obtained from our experiments provide valuable insights into the effectiveness of different methods for predicting the country of origin of Asian Last.fm users based on their music preferences.

The Node2vec embeddings combined with Catboost achieved the highest accuracy of 87.2%, closely followed by the Node2vec embeddings with LGBM at 87.0%. The Node2vec embeddings combined with XGBoost also yielded a respectable accuracy of 86.6%. These results confirm the suitability of graph-based machine learning techniques for analyzing social network data and leveraging the connections between users and their music preferences to infer their country of origin.

However, the SVD-based methods exhibited lower accuracy scores. SVD with LGBM achieved an accuracy of 70.1%, followed by SVD with XGBoost at 78.3% and SVD with Catboost at 61.6%. When we concatenated adjacency matrix with feature matrix and retrained models, the quality decreased by almost 10%. These results indicate that the SVD approach may not fully capture the complex relational information and structural patterns present in the Last.fm social network, resulting in lower predictive performance compared to the Node2vec embeddings.

The Graph Attention Neural Network (GAT) achieved an impressive accuracy of 90.6%, surpassing all other methods. This result emphasizes the power of deep learning models specifically designed for graph-structured data in capturing intricate relationships and patterns within social networks. The GAT model effectively leveraged the graph's structure and attention mechanisms to learn meaningful node embeddings, leading to highly accurate country predictions.

4 Conclusion

In conclusion, the LastFM Users Country Prediction project aimed to predict the country of origin of Asian Last.fm users based on their music preferences. Through our analysis, we compared several methods and identified the most effective approach for this task.

Among the methods tested, the Graph Attention Neural Network (GAT) emerged as the top-performing model, achieving an impressive accuracy of 90.6%. The GAT model, designed specifically for graph-structured data, successfully captured the complex relationships within the Last.fm social network, leveraging attention mechanisms to learn meaningful node embeddings. This high accuracy demonstrates the power of deep learning models in accurately predicting users' country labels based on their music preferences.

The Node2vec embeddings combined with Catboost also demonstrated strong performance, achieving an accuracy of 87.2%. This approach effectively captured the graph's structural information and relationships between users, highlighting the role of music preferences in inferring cultural backgrounds. On the other hand, the SVD-based methods exhibited lower accuracy scores, suggesting that they may not fully capture the intricate relationships present in the Last.fm social network.

Overall, the results affirm the significance of graph-based machine learning techniques in understanding cultural differences and music preferences within the Asian Last.fm user community. The successful implementation of the GAT model and the notable accuracy achieved by the Node2vec embeddings with Catboost contribute to a deeper understanding of the best approaches to tackle node classification problem.

5 References

References

1. B. Rozemberczki and R. Sarkar, "Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models," in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, p. 1325–1334, ACM, 2020.