Data Mining Report

Beatriz Moreira
Faculdade de Ciências
Universidade de Lisboa
fc54514@alunos.fc.ul.pt
Hours: 8H

João Lobato
Faculdade de Ciências
Universidade de Lisboa
fc62611@alunos.fc.ul.pt
Hours: aa

Rute Patuleia
Faculdade de Ciências
Universidade de Lisboa
fc51780@alunos.fc.ul.pt
Hours: 8H

Tiago Assis
Faculdade de Ciências
Universidade de Lisboa
fc62609@alunos.fc.ul.pt
Hours: aa

1. Introduction

The present project aims to predict the interaction activity of molecules based on a dataset containing data for 144 G protein-coupled receptor (GPCR) annotated proteins. The target variable for this model is the molecular activity, with value ratings ranging from 1 to 10, where 1 means inactive and 10 means extremely potent. The molecules were identified by their ChEMBL IDs and the proteins by their Uniprot IDs. Predicting molecular activity is crucial in the field of drug discovery, as it helps in identifying potential candidates for therapeutic use by assessing how strongly they interact with target proteins.

2. Objective

The main goal of this project was to accurately predict the interaction activity values of each protein-molecule pair present in the activity_test_blanked.csv file. These predictions are based on known interactions described in the activity_train.csv file. The aim is to fill the zeroed activity values in the test dataset with accurate predictions ranging from 1 to 10, thereby demonstrating the effectiveness of the predictive model.

3. Data Description

Regarding the train set, this contains interactions between molecules and proteins with activity values ranging from 1 to 10. The test set displays a similar structure to the previous one but with activity values set to zero, which need to be predicted. Finally, the mol_bits.pkl contains hashed structural representations (fingerprints) of molecules, with each set bit representing a specific structural feature of the molecule out of a total of 2048 features.

The dataset encompasses a total of 140339 interactions, with 135711 interactions used for training and 4628 for validation, involving 144 protein targets and 73865 different molecules.

4. Preprocessing

The molecular fingerprints provided in the mol_bits.pkl were processed to create vectors for each molecule. These features were processed by feature vectorization resulting in 2048-dimensional vectors for each molecule, where each position corresponds to a potential structural feature, with set bits indicating the presence of those features. Additionally, each unique protein and molecule categorical identifier was mapped to an integer value which functioned as an index to facilitate array lookups and embeddings.

5. Technical Approaches

5.1. Collaborative Filtering

Collaborative Filtering (CF) is the process of making recommendations of items to users based on the behavior of other similar users. In the case of this dataset, the proteins are akin to users and molecules are akin to items, with the activity values being the "rating" of each user (protein) for each item (molecule). This approach assumes that if certain proteins interact similarly with various molecules, they might show similar interaction patterns with other molecules as well, i.e., the goal is to predict the activity of an unknown protein-molecule pair based on other similar interactions where the protein and the molecule are also present. User-User Collaborative Filtering (UU-CF)

is a type of CF that can be used to infer the activity of each new protein-molecule pair by computing similarities between proteins based on their existing interactions with molecules.

The process starts by transforming the dataset into a user-item matrix, where each row is a protein and each column is a molecule. The cell values are the activity values filled with nulls where no activity previously exists. Then, each row of the matrix is centered by subtracting its mean value. Effectively, the null values become zeros (the mean), and every other value is transformed into the offset from the mean. This normalizes the scores and helps in comparing different proteins that might have different activity baselines, removing some of the individual biases. Then, the user-user similarity matrix was calculated, using the cosine similarity allowing us to identify proteins that have similar interaction patterns. Finally, the Global Baseline Average (GBA) estimator was also computed by finding the global activity average, the average activity for each protein, and the average activity for each molecule. These values are stored in individual matrices and are used to inform predictions by incorporating general trends in the dataset, and to be used to predict pairs with no previous information. This value is defined for user u and item i as:

$$GBA_{u,i} = M + D_i + D_u \tag{1}$$

where M is the global activity average, D_i is the offset between M and the activity average for item i, and D_u is the offset between M and the activity average for user u. The estimated activity score of a molecule i with a protein x is then calculated by selecting the top k nearest neighbors (most similar) and then calculating the respective weighted average taking into account the GBAs:

$$r_{xi} = GBA_{xi} + \frac{\sum_{j \in N(i;x)} s_{ij}(r_{ij} - GBA_{xj})}{\sum_{j \in N(i;x)} s_{ij}} \tag{2}$$

where $N(i;x)$ are the nearest neighbors of molecule i who interact with protein x, r_{xj} is the activity score of a molecule j with protein x, GBA_{xi} is the global baseline average for molecule i and protein x, and GBA_{xj} is the global baseline average for molecule j and protein x. For this work, a k value of 10 was chosen.

5.2. Matrix Factorization

Matrix factorization techniques like SVD decompose large matrices, which are typically very sparse, into products of smaller matrices:

$$R = PQ^T \tag{3}$$

The reconstruction of the original matrix by the dot product of these new matrices can be used to train the decomposed matrices in a way that eliminates sparsity and optimizes the latent factors to allow the prediction of unobserved relations. Latent factors represent underlying dimensions that try to explain the observed interactions. Training is performed using Stochastic Gradient Descent (SGD) to minimize the regularized least squares equation:

$$\min_{P,Q} \sum_{u,i} (r_{ui} - p_u q_{iT})^2 - \lambda \left(\sum_u \parallel p_u \parallel + \sum_i \parallel q_i \parallel \right) \tag{4}$$

where r_{ui} is the activity score for protein u and molecule i, p_u is the latent vector for protein u, q_i is the latent vector for molecule i, and λ is the regularization term.

The best hyperparameters were found using grid search with 5-fold cross-validation, thus training was performed during 25 epochs with 25 latent factors, a learning rate of 0.01, and a λ of 0.1.

5.3. Neural Collaborative Filtering

Neural Collaborative Filtering (NCF) is a modern approach to recommendation systems that utilizes deep learning techniques to predict user interactions with items. Unlike matrix factorization methods, which try to optimize a linear multiplication between two latent factor matrices, NCF employs neural network architectures to model complex and non-linear relationships inherent to the interactions. [1]

In this approach, each protein and molecule IDs were mapped to indices that are used as input to the neural network. The neural network then converts these indices into dense continuous embedding vectors of each protein and molecule which functions as a compact representation. These embedding vectors are learned during training to capture latent factors. After the embedding layers, the molecule structural features are concatenated with both of the embedding vectors and are passed through fully connected hidden layers to learn and capture complex interactions between the proteins and molecules with the auxiliary aid of the hashed molecule structural representations. Each hidden layer was followed by a ReLU activation function. Additionally, dropout was performed after each hidden layer to reduce overfitting. The final output layer predicts the activity value of a protein-molecule pair.

Several hyperparameter combinations were tested, and the best results were obtained by training during 100 epochs with a learning rate of 5×10^{-5}, a batch size of 128, an embedding dimension size of 32, an AdamW optimizer with 10^{-5} weight decay, a cosine annealing learning rate scheduler, and the MSE loss. Three hidden layers were employed with dimensions of 1024, 256, and 64, respectively, and a neuron dropout probability of 0.5. Models were trained in a single Nvidia Tesla P100 with 16GB VRAM.

Table 1. Model performance results obtained for the validation set. In bold and underlined are the best and second-best values for every metric, respectively.

Model	RMSE	MAE
NCF	<u>2.4364</u>	**1.9143**
SVD-MF	**2.4225**	<u>1.9898</u>
UU-CF	2.6415	2.0389

5.4. Model Evaluation

The models were evaluated on 80% training and 20% validation splits to fine-tune hyperparameters and ensure the robustness of predictions. During evaluation, the training split was used to train the models, and the validation split was used to evaluate the predictions. The metrics used were the root mean squared error (RMSE) and mean absolute error (MAE), from which we can interpret the magnitude of the average difference between the predicted and the actual activity values.

For the UU-CF model during evaluation, the activity values for protein-molecule pairs used for validation were set to null in the interaction matrix to simulate how this model would behave in predicting unknown interactions.

Besides the grid search performed for SVD-MF, no cross-validation was performed while evaluating the NCF and UU-CF models. This is a limitation of this work.

5.5. Final Predictions

The final predictions for the test dataset were generated using an ensemble approach by combining the outputs of the NCF, UU-CF, and SVD-MF models trained with the full training dataset. The ensemble was simply the average of the three obtained predictions. The final activity predictions should now be more robust than those obtained by the three models separately. The predicted activity values were clipped to the range from 1 to 10 to match the target range.

5.6. Results

????????????????? NEW VALUES ????????????????? ????????????????? NEED TO CHANGE ????????????? The NCF model, leveraging embeddings for proteins and molecules, with an additional molecular fingerprint feature vector and non-linear relationship modeling, yielded an RMSE of 2.4364 and an MAE of 1.9143 on the validation set, indicating a moderate level of prediction accuracy. Additionally, SVD-MF achieved similar results with an RMSE of 2.4225 and an MAE of 1.9898 on the validation set, indicating a similar performance to the NCF model. This is expected as they share key concepts involving matrix operations and latent factor optimization. The fact that SVD-MF achieves a better RMSE than NCF with a lower MAE indicates that, although its average error magnitudes are smaller, the overall deviation from the ground truth is larger. On the other hand, NCF has more occasional large errors (thus a higher RMSE), which may be a result of possibly overfitting some points.

The UU-CF model showed worsened performance with an RMSE of 2.6415 and an MAE of 2.0389, showing the downsides of this method, such as data sparsity and cold start. In cases where a protein has only a few interactions with molecules, finding a sufficient number of similar proteins becomes challenging and leads to less accurate results, contrary to methods that focus on learning latent factors that can deal better with missing information.

The predictions from these models were combined through an ensemble approach with the expectation that these results would lead to even more robust and reliable predictions for the test dataset with unseen interactions. The ensemble approach harnesses the strengths of each individual model, yielding predictions that are more accurate than those from any single model alone.

6. Conclusion

This study showed that using Collaborative Filtering is a good tool for predicting the activity of protein-molecule pairs. The use of three types of recommendation-based models allowed for a comparison between them, in which the best-performing approach was the UU-CF.

The combination of the models in an ensemble was utilized to improve the accuracy and robustness of the obtained activity predictions.

This problem could be approached differently by considering this a classification task and using the target labels as different classes to predict. However, this would be suboptimal as the predictions would be limited to finite whole numbers resulting in larger errors and lower prediction expressiveness.

References

[1] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, page 173–182, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. 2