Prospeção de Dados
Universidade de Lisboa - FCUL
07 de junho de 2024

Course Project

Jorge Aleluia (MI - 54549) - 6 hours
Rómulo Nogueira (MEI - 56935) - 11 hours
Guilherme Cepeda (MEI - 62931) - 8 hours
Catherine Prokhorov (MCD - 62608) - 8 hours

The code, report, and predictions are available online through a GitHub repository: `https://github.com/RomuloNogueira02/DataMining/blob/main/Project`

1 Objectives

The main objective of this project is to predict the activity values between proteins and molecules based on data on molecular interactions and structural fingerprints. Specifically, the task involves using the provided training dataset *activity_train.csv*, which contains a lot of protein - molecule pairs and the structural fingerprint data *mol_bits.pkl* to predict the activity values for the pairs protein - molecules listed in *activity_test_blanked.csv*.

2 Data Analysis

Before starting, we analyzed the data (using the Data Wrangler tool and different notebooks) through a few steps, to gain some insight into the data and understand the best approach for making predictions.

We loaded the molecular fingerprints from *mol_bits.pkl*, which contain hashed structural representations of molecules, and the training and test datasets from *activity_train.csv* and *activity_test_blanked.csv*, respectively.

After loading the data, we summarized the dataset to verify the number of unique molecules and the number of rows in the training and test datasets. Specifically, we had 73,865 unique molecules, 135,711 rows in the training dataset, and 4,628 rows in the test dataset.

3 Pre-processing Data

3.1 Cleaning and Formatting

In the data pre-processing stage, we focused on cleaning and correctly formatting the data, by removing blank spaces from the molecular IDs and loading the molecular fingerprints. This step was crucial to ensure the data was consistent and ready for analysis and modeling.

Then, we randomly selected a validation subset from the training data, consisting of approximately 1/3 of the total training lines for further analysis and model evaluation.

3.2 Creating MinHashLSH Structure

To efficiently identify similar molecules based on their structural features, we created a **MinHashLSH** (Locality Sensitive Hashing) structure using the *datasketch* library that provides us with the necessary methods to do it.

For each molecule, a MinHash signature was generated, with 256 permutations, to approximate Jaccard similarity. This involves hashing each element of a set multiple times with different hash functions and taking the minimum value for each function, forming the MinHash signature. The generated MinHash signatures were then inserted into an LSH structure.

LSH hashes similar items into the same buckets with high probability, facilitating efficient searching of similar molecules. The LSH structure was queried using Jaccard similarity to find molecules similar to a target molecule.

This allowed us to quickly retrieve a set of structurally similar molecules, significantly reducing computational complexity.

3.3 Activity Matrix

We created an activity matrix where rows represent proteins and columns represent molecules, with each cell containing the activity level of a particular molecule-protein pair.

Centering the matrix involved subtracting the mean activity level of each row from all entries in that row. This normalization step adjusts for different baseline activity levels across proteins, ensuring that similarity calculations are more meaningful.
Mathematically, for each entry a_{ij} in the activity matrix A, the centered value $\tilde{a}_{ij}$ is computed as:

$$\tilde{a}_{ij} = a_{ij} - \bar{a}_i$$

where $\bar{a}_i$ is the mean activity level of protein i (i.e., the mean of the i-th row).

Since the activity matrix can have missing values, the mean for each row was calculated ignoring the missing values, and these missing entries were filled with the centered values appropriately. Centering the activity matrix improves the accuracy of similarity calculations, ensuring that similarity scores are not skewed by different activity baselines among proteins. This leads to more reliable identification of similar proteins and better prediction performance in the collaborative filtering algorithm.

4 Hybrid Approach

We decided to implement a hybrid recommender system that integrates user and item-based methods, which are well-known in collaborative filtering.

In the context of our project:

- **User-Based Methods:** These methods focus on comparing users to make recommendations. In our scenario, we treated the proteins as user analogs, where this

part of our system can recommend the most similar proteins to a target protein based on the level of activity of the proteins with the molecules.

- **Item-Based Methods:** These methods compare items to each other. Once we have the proteins that are similar to the target protein, our recommender system uses the proteins' molecules that are similar to each other to estimate the level of activity.

Given an input pair *(protein, molecule)* our final pipeline looks like this:

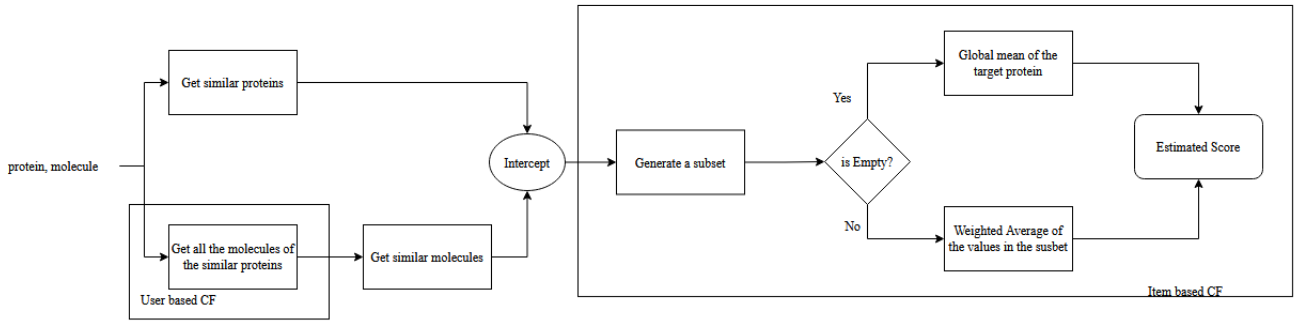

Figure 1: Full pipeline

4.1 User-based Collaborative Filtering

To obtain similar proteins, we used the centered matrix (obtained in the data preprocessing layer) and calculated the Pearson Correlation Coefficient with all the other proteins in the training data set. This resulted in an Mx1 matrix where each entry represents the similarity.

This matrix is then filtered by a minimum threshold, and the 5 most similar proteins (including the input protein), along with their similarity scores, are returned (if there are 5).

4.2 Intermediate steps

Next, using the MinHashLSH Structure created in the preprocessing layer, we obtain the K-NN molecules closest to the one given as input, also filtering by a minimum threshold.

We also retrieve all the molecules from the similar proteins identified earlier, enabling us to find the intersection between these molecules and those similar to the input molecule.

This provides us with information on the activity of similar molecules within similar proteins, which serves as a solid foundation for calculating the final score.

4.3 Item-based Collaborative Filtering

As the final step, we use the item-item Collaborative Filtering method.

First, we check for any intersections. If no intersection exists, we use the Global Baseline Average to estimate the score based on the activity of the input protein with all its molecules.

If an intersection does exist, meaning there are similar proteins that share similar molecules, we proceed with the calculation based on a weighted average. This step is conducted in two main stages.

In the first stage, we iterate through each set of molecules for each protein, storing the value of the numerator (the sum of the product of molecule similarity level and activity) and the denominator (the sum of molecule similarity levels).

Next, we iterate through each protein and calculate the score by dividing the final numerator by the denominator.

This first stage essentially employs the following formula to obtain an estimate from each similar protein:

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

s_{ij}... similarity of items i and j
r_{xj}...rating of user x on item j
N(i;x)... set of items rated by user x similar to item i

Figure 2: Used formula

Finally, in the second stage, we estimate the final value by weighing the average of the activity level by the similarity score of each protein to the input protein.

5 Evaluation

5.1 Model Testing and Validation

The hybrid collaborative filtering model was tested on the validation set. Predictions were compared with actual activity values, and the performance was evaluated by plotting and analyzing different predictions against the actual values.

The error percentage for each prediction was calculated to evaluate model accuracy and it was analyzed to categorize predictions based on their accuracy levels:

- $error <= 25\%$ - is considered good performance;
- $25\% < error <= 50\%$ - is considered moderate error;
- $error > 50\%$ - is considered high error;

This categorization provided insights into the model's performance and highlighted areas for potential improvement. A scatter plot visualizing the relationship between predicted and actual activity values was created. This plot helped assess how well the model's predictions aligned with the true activity levels.

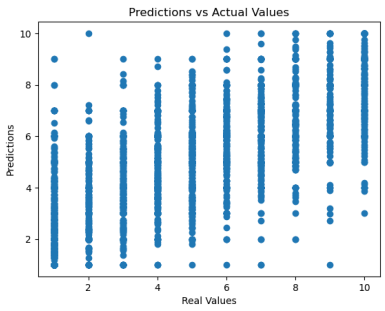

Figure 3: Predictions Vs Actual Values

		abs			%		
		[0, 25]	]25 ,50]	]50, +inf]	[0, 25]	]25 ,50]	]50, +inf]
knn prot	0,25 0,25	2056	1099	957	50%	26.70%	23.30%
knn prot	0,5 0,25	2153	1047	912	52.40%	25.50%	22.20%
knn prot	0,75 0,25	1966	935	1211	47.80%	22.70%	29.50%
knn prot	0,25 0,5	2242	1057	813	54.50%	25.70%	19.80%
knn prot	0,25 0,75	2433	1009	670	59.20%	24.50%	16.30%
knn prot	0,5 0,5	2317	1035	760	56.30%	25.20%	18.50%
knn prot	0,5 0,75	2505	970	637	60.90%	23.60%	15.50%
knn prot	0,75 0,5	2090	893	1129	50.80%	21.70%	27.50%
knn prot	0,75 0,75	2203	854	1055	53.60%	20.80%	25.70%

Figure 4: Preditions Results

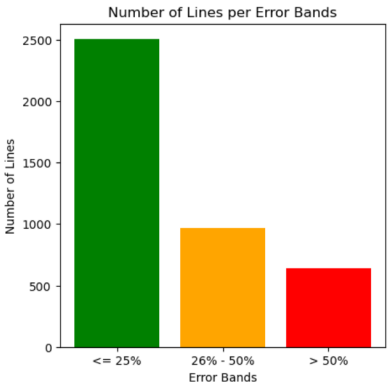

Figure 5: Nº of lines per Error Bands

We tested different thresholds for the similarity level of proteins (THRESHOLD_PROTS) and molecules (THRESHOLD_MOLS), obtaining the best results with these set at 0.75 and 0.5 respectively.

The table shows that the model used to predict the level of activity between proteins and molecules is very effective. Most of the interactions, representing approximately 2500 lines, have an error of less than 25%, indicating high accuracy in the model's predictions.

There is a smaller but still significant, number of predictions with moderate errors, between 26% and 50%, with around 1000 lines. This shows that although the model is not perfect, it still provides reasonably accurate results for many interactions.

Only a small portion of the predictions, approximately 500 lines, have high errors of over 50%. These are the areas where the model faces the most difficulties and where there is the greatest need for improvement.

In summary, the data indicates that the model is effective in most cases, but there is room for adjustment, especially in the most challenging predictions. Analyzing these areas of high error can provide valuable insights for improving the model's accuracy.

6 Conclusion

In this project, the main objective was to predict the activity values between proteins and molecules based on data on molecular interactions and structural fingerprints. We used a training dataset and created a hybrid collaborative filtering structure to improve the accuracy of the predictions. The hybrid approach combined user-based and item-based techniques, utilizing the MinHashLSH structure to identify similar molecules and Pearson Correlation Coefficient calculations to find analogous proteins.

Even though we don't have the ground truth of the test dataset, we believe that our approach (and this is based on validation) manages to produce interesting results that are close to the real activity levels, due to a hybrid implementation that allowed us to use data from different sources, thus enriching the final calculation of the estimate.