Dataset:

Collaborative filtering approaches are usually designed to work on very large datasets. Therefore, it is crucial that the proposed approach is scalable. To find out the best scalable approach and give a more real-life scenario with millions of users, we’ve used a huge dataset open-sourced by eHarmony, an online dating platform. The dataset provides user to user matching data to show whether two individuals; given by their respective user IDs matched or not which basically shows their mutual interest. This has been indicated by positive and negative examples in the dataset wherein positive examples have been depicted by one and negative examples as zero. Two users have been considered to be relevant or similar if they match, otherwise, they have been regarded as irrelevant. The users have been kept anonymous to ensure their privacy and therefore numbers have been used instead. Similarly, to describe an individual, his personal choices or profile details have been written in numerical form. The mapping has not been explained by the authors. Every user has been represented as a vector with 59 columns, one of them is their User\_ID and the other 58 columns have been used to describe their profile including their interests and their personality.

The dataset has been divided according to the matching time of the two users. Training data comprises of the matches of the users in the first half of the time interval and testing data contains the matches in the other half. The training dataset consists of approximately 2,75,000 unique users and approximately half a million matches from the eHarmony app. Such a huge dataset is difficult not just to process but might also contain some users which are not contributing to the discriminative data. Therefore, we sample the data, to get the top --- users with ---.

Algorithm:

We are trying to test the prediction methods used as the state of the art for drug-target interaction and compound-protein interaction.

DTI\_Predict, Drug Target Interaction Prediction algorithm

Our baselines are(refer to surprise package for all):

SVD

Negative Matrix Factorisation

Normal Predictor

References:

<https://bmcfee.github.io/papers/mlr.pdf>

<http://admis.fudan.edu.cn/projects/pucpi.html>