Chapter 4

Method of Evaluation

In this section, we discuss the evaluation methods to determine the quality of our proposed recommendation methods.

**Experiment**

After building the model and applying the methods on our training dataset, we apply the methods on our test dataset for comparing the accuracy. From the dataset of over 1.9 million observations of projects and around 1 million users, the data set was divided into 80% as training data and 20% was used as test data for the model.

**Evaluation methods**

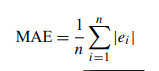
The area of recommender system research has proposed a number of statistical measures to evaluate the quality of a recommender system. These measures can be broadly categorized into two groups.

**Statical accuracy metrics**

A Statistical accuracy measure that accesses the accuracy of a recommender system by finding the difference between the recommendation score and the actual user score for the user-item combination in the test dataset.

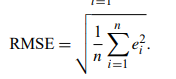
**MAE(Mean Absolute Error)** is one of the most popular metrics that is widely used to compare the actual and expected values. Basically, MAE is the measure of the deviation of the predicted recommendation rating from the actual user given ratings. In our dataset for each pair of {predicted, user-rating }, MAE handles the absolute difference between them equally. I.e. it is given by | predicted - user rating |. MAE is computed by summing the absolute difference for all the {predicted, user-rating} pairs and then averaging it over the total number of pairs. Mathematically it is given by,

MAE = SUM(predicted - user-ratings)/Total number of pairs



The lower the MAE value, we say that accurate is the prediction score. And thus the more accurate is our recommendation system.

**Root Mean Squared Error (RMSE)** is another statical accuracy matric that is used for measuring model performance. It was first used in the areas of environmental changes and climate research [46]. While MAE assigns the same weight to all the errors, RMSE penalizes large variance in the reading as it assigns errors with large absolute value more weight compared to the smaller values.



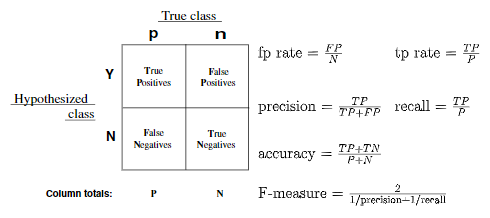
**Correlation** is another statical accuracy metrics that are used to compare and evaluate the results of the model.

**Decision support accuracy metric**

Decision support accuracy metrics are used to measure the usefulness of a recommendation engine at assisting the user in selecting the top-quality items from a list of all available items. They work by assuming the predictions to be a binary process. Either an item is a good one or a bad one, i.e. it is either predicted or excluded from prediction. If on a rating scale of 5 if a user only chooses items with a rating more than 4, it becomes irreverent if the rating is 2.5 or 3. Some of the most widely utilised support accuracy matrices are ROC sensitivity, weighted error and reversal rate [37].

**ROC sensitivity - Receiver operating characteristics curve** is a graphical plot that is used to visualise and select classifiers based on their performance. ROC graphs have been used for a long time in the signal detection system for showing variance in a hit and false rates of the classifiers [47]. The medical community has a huge literature on the use of ROC for decision making on diagnostic testing.ROC curves are adopted in machine learning for comparing and evaluating various algorithms. ROC graphs have recently gained popularity in the machine learning community due to the understanding that simple accuracy metrics are a weak measure of the classification performance [48]. Furthermore, of being useful as a performance mapping technique, they have additional qualities that make them suitable for problems with very skewed data distribution.

We begin with building a confusion matrix for the classifier, counting the number of correctly classified and incorrectly classified values into the matrix. As shown in figure xx the true positive and true negatives are the correctly classified values by the classifier and false positive and false negative are the incorrectly classified values.

****

**FIG: performance metrics calculations from Confusion matrics.**

Figure xx shows various standard metrics that can be calculated from the confusion matrix. Formally it is defined as the area under the curve of the ROC curve. ROC is plotted as sensitivity v/s (1- specificity) fo the test. In case of a recommender system Sensitivity is defined as the probability of a random selection of item is accepted by the filtering method (a good prediction) . and specificity is defined as the probability of a random selection of item being rejected by the filtering method (a bad prediction). A good filtering method would enable the user to select from 90% of the good recommendation and 10% of the bad ones.

**Reversal rate** is defined as the frequency at which the recommendation by the system is wrong or incorrect. On a scale of 5, it is commonly given by the percentage of the recommendation where the prediction was varied by 3 or more points.

**Weighted error** is the metrics in which gives extra weight to the errors due to the strong outlook about the recommended item. For instance, if a user considers an item to be his favourite (i.e. a rating of 5 out of 5) the error for this could be considered double or more.

A more novel way is proposed in [24] based on the coverage and Serendipity of the recommendations. This discusses the matrics beyond the accuracy of the recommendations and takes into consideration the quality and usefulness of the system.

For the simplicity and ease of interpretation of MAE, we select it as our preferred choice of evaluation. In the experiment [37] \Sarwar notes that MAE and ROC give the same arrangement of different experimental procedures with respect to prediction quality.

**“Results**

In this section, we discuss the results of applying the different item-item collaborative filtering algorithms on the Scratch dataset.

**Comparing the similarity algorithms**

After running the three similarity computation methods namely cosine, adjusted cosine and pearson-r correlation as shown in previous chapters on our training data and using the weighted sum method to get the predictions score, we calculate the MAE of for each of the methods. The figure shows the results of the three algorithms on the dataset.

