Abstract

zMovie is a personal movie recommendation service that gives custom movie suggestions based on previous user ratings by comparing them to those of similar users in the zMovie database. The core engine is based on a user-model based collaborative filtering algorithm using clustering.

The Algorithm

STEP 1: Randomly select $K$ users as centroids (mean-point) of $K$-clusters
STEP 2: Run $P$ passes over “base” dataset to cluster all users based on similarity function
STEP 3: Each new user rates at least $M$ movies
STEP 4: Assign new user to the most similar cluster
STEP 5: Use the $N$ most similar users from the cluster to provide recommendations for the user

Analysis

Mean Absolute Error (MAE)
Sum of all absolute errors (predicted rating – actual rating) for all items in test set divided by size of test set.

The effectiveness of the algorithm is measured by calculating MAE. Thus, the combination of parameters are chosen to minimize MAE.

Choosing Parameters
- $K$ – Too few clusters results in grouping dissimilar users whereas too many clusters defeat the purpose of grouping “like” users.
- $P$ – After a certain number of passes, each additional pass of the $K$-means algorithm gives lower incremental improvement in clustering of dataset.
- $M$ – Working with too few initial ratings makes it harder to classify the new user into an existing cluster.
- $N$ – Using too few similar users would not be representative of the cluster and using too many would defeat the purpose of the optimization.

Conclusion

Most Optimal Values (yielding lowest MAE)
$K = 10 \quad P = 7 \quad M = 10 \quad N = 20$

- User-model based collaborative filtering algorithms, using clustering, can form a basis for good recommendation systems
- Choosing parameters of the algorithm optimally have a big impact on performance
- Using additional metadata such as content based factors like genre may improve performance

Dataset

- **MovieLens Dataset**
  - 100,000 ratings (scale of 1-5)
  - 1,682 movies
  - 943 users

Global dataset split: 50% users in “base” dataset; 50% users in “test” dataset