Recommendation Systems

Machine Learning Final Project

Arezoo Rajabi
Introduction

- Increasing spread of the Internet → appearance of business and trade opportunities
- Popular among these businesses = E-Shopping
- Some Of Well-known Commercial Systems
  - Amazon
  - Movielens (movie recommender system)
Recommendation System

- Goal: estimating the users’ interest in items that they have not seen yet
- The forecast operation is done according to the users’ and items’ information or the ratings of items assigned by the users

\[ R: \text{User} \times \text{Item} \times \text{Context} \rightarrow \text{Rating} \]
Dataset

- Input: Matrix of Users and Their Ratings
- Features:
  - So sparse
  - Integral Ratings
  - High dimensionality
Common Methods[5]

• Computing Similarity between users
  – based on similarity of their ratings to items
  – Find similar users to the target user and predicting the amount target user's interest in unranked items

• Computing Similarity between Items
  – Based on Similar rates that are given to them
  – Find similar Items to items that the target user was interested in to propose
Proposed Method

- Finding similar users or Items plays an important role in Recommendation System
- Sparsity is one of the main problem in these Systems
- Proposed Method:
  - Combining Features into Some Groups
Selected Methods

- Working well with non-numeric data
- Fast in building model
- Chosen Methods
  - M5P
  - Random Forrest
  - Random Tree
  - Decision Table
M5P[2]

- Trees of regression models
- A decision-tree induction algorithm is used to build a tree
- Splitting criterion: minimizing the intra-subset variation in the class values down each branch
- M5P stops if the class values of all instances that reach a node vary very slightly, or only a few instances remain.
REPTree[6]

- A fast decision tree learner
- Using information gain as the splitting criterion
- Prunes it using reduced error pruning
Random Forest Tree[6]

- Ensemble of unpruned classification or regression trees
- Induced from bootstrap samples of the training data
- Using random feature selection in the tree induction process
- Prediction is made by aggregating
**Decision Table[3]**

*Decision tables* are a precise yet compact way to model complicated logic

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Printer does not print</td>
<td>Y Y Y Y N N N N</td>
</tr>
<tr>
<td>A red light is flashing</td>
<td>Y Y N N Y Y N N</td>
</tr>
<tr>
<td>Printer is unrecognized</td>
<td>Y N Y N Y N Y N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check the power cable</td>
<td>X</td>
</tr>
<tr>
<td>Check the printer-computer cable</td>
<td>X</td>
</tr>
<tr>
<td>Ensure printer software is installed</td>
<td>X X</td>
</tr>
<tr>
<td>Check/replace ink</td>
<td>X X</td>
</tr>
<tr>
<td>Check for paper jam</td>
<td>X X</td>
</tr>
</tbody>
</table>
Dataset

- Movie
  - Movie ID
  - Movie Name
  - Genre

- User
  - User ID
  - Gender
  - Occupation
  - Age

- Rating
  - (User ID, Movie ID, Rate)
Defects of Dataset

• Sparsity: Only 200,000 ratings for 6040 users and 1600 movies
• High amount of low rated movies
• So big for common machine softwares (Weka)
New Dataset

- (User ID, Age, Occupation, Gender, Genre, Genre Average)
- Low Dimension Data
- Using Average of Genre ratings instead of Movies as Item
- Less sparsity
- Losing part of data
## Result

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation coefficient</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Relative absolute error %</th>
<th>Root relative squared error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>M5P</td>
<td>0.2336</td>
<td>0.5209</td>
<td>0.6886</td>
<td>96.3103</td>
<td>97.2419</td>
</tr>
<tr>
<td>REPTree</td>
<td>0.1815</td>
<td>0.5331</td>
<td>0.7043</td>
<td>98.5676</td>
<td>99.4541</td>
</tr>
<tr>
<td>Random Forest Tree</td>
<td>0.1082</td>
<td>0.6144</td>
<td>0.806</td>
<td>113.599</td>
<td>113.828</td>
</tr>
<tr>
<td>Decision Table</td>
<td>0.2347</td>
<td>0.5207</td>
<td>0.6885</td>
<td>0.806</td>
<td>97.2232</td>
</tr>
</tbody>
</table>
New Dataset

- (UserID, Age, Occupation, Gender, Genre1, Genre2,...)
- Adding a feature for each Genre
- Assigning zero value to Genres that users have not rated
## Different Algorithm on Action Genre

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correlation coefficient</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Relative absolute error</th>
<th>Root relative squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>M5P</td>
<td>0.7287</td>
<td>0.2455</td>
<td>0.4384</td>
<td>52.9649</td>
<td>68.5294</td>
</tr>
<tr>
<td>REPTree</td>
<td>0.6944</td>
<td>0.2759</td>
<td>0.4612</td>
<td>59.5364</td>
<td>72.0889</td>
</tr>
<tr>
<td>Random Forest Tree</td>
<td>0.721</td>
<td>0.2544</td>
<td>0.4434</td>
<td>54.8865</td>
<td>69.3056</td>
</tr>
<tr>
<td>Decision Table</td>
<td>0.6623</td>
<td>0.2847</td>
<td>0.4799</td>
<td>61.4267</td>
<td>75.0133</td>
</tr>
</tbody>
</table>
M5P for different Genres

<table>
<thead>
<tr>
<th>Genre</th>
<th>Correlation coefficient</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Relative absolute error</th>
<th>Root relative squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>0.7287</td>
<td>0.2455</td>
<td>0.4384</td>
<td>52.9649</td>
<td>68.5294</td>
</tr>
<tr>
<td>Documentary</td>
<td>0.818</td>
<td>0.6394</td>
<td>1.1079</td>
<td>35.5224</td>
<td>57.4763</td>
</tr>
<tr>
<td>Crime</td>
<td>0.5876</td>
<td>0.5333</td>
<td>0.9067</td>
<td>71.1203</td>
<td>80.9141</td>
</tr>
<tr>
<td>Comedy</td>
<td>0.671</td>
<td>0.2598</td>
<td>0.3791</td>
<td>66.9328</td>
<td>74.3845</td>
</tr>
<tr>
<td>Children</td>
<td>0.666</td>
<td>0.6663</td>
<td>1.0289</td>
<td>64.5464</td>
<td>74.6084</td>
</tr>
<tr>
<td>Animation</td>
<td>0.6138</td>
<td>0.8923</td>
<td>1.2919</td>
<td>66.7254</td>
<td>78.9787</td>
</tr>
<tr>
<td>Advanture</td>
<td>0.6295</td>
<td>0.3421</td>
<td>0.6113</td>
<td>62.4949</td>
<td>78.2095</td>
</tr>
<tr>
<td>Drama</td>
<td>0.9932</td>
<td>0.0155</td>
<td>0.0926</td>
<td>2.5505</td>
<td>11.6543</td>
</tr>
<tr>
<td>Romance</td>
<td>0.5286</td>
<td>0.3351</td>
<td>0.5677</td>
<td>72.4879</td>
<td>85.219</td>
</tr>
<tr>
<td>Sci Fi</td>
<td>0.6394</td>
<td>0.342</td>
<td>0.6152</td>
<td>61.9196</td>
<td>76.8941</td>
</tr>
</tbody>
</table>
RRSE & RAE relation with CC

Correlation Coefficient

RRSE

Correlation Coefficient

RAE
MAE & RMAE relation with CC

Correlation coefficient
Documentary and Drama Distribution
References