Improve result of prediction using user information

Seminar Report

An. La

Data scientist - Big Data team
FPT Telecom, HCM Vietnam

24 Feb, 2018

Outline

The Method - Motivation of ANN

- Adaptive for various types of problems.
- Flexible adjusting to fit best with problems.
- ...

The Method - Preliminary I

1. Data pre-propressing
   - Remove unused variables (zipcode, timestamp).
   - Convert categorical variable into indicator variables (genres, gender, occupation).
   - Treat age as continuous variable.
2. Training set - validating set - testing set
   - Separate list of users to user_train, user_val, user_test with ratio: 0.64:0.16:0.2
   - Create training/validating/testing set from mapping ratings data with corresponding user list.
     training set:validating set:testing set = 0.63:0.17:0.2
3. Scaling
   - Determine min and max through training set.
   - Apply min-max scale on validating and testing set.

The Method - Preliminary II

- Evaluation
  Calculate Root Mean Square Error (RMSE) on all items:

$$r = \sqrt{\frac{1}{n}\sum_{i}^{n}(predict - target)^2} \quad (1)$$

Relative Change between 2 results (RC):

$$rc = \frac{r - r_{ref}}{r_{ref}} * 100\% \quad (2)$$

The Method - Preliminary III

- Definitions of ANN
  - Sigmoid function
  - Fully Connected
  - Batch Normalization
  - Early Stopping

Model Architecture

- Baseline and No user-info model: 3 layers (item-info vector (input), hidden, average ratings (output)).
- Stacking user-info model: Concatenate item-info vector and user-info vector. See figure.
- Embedding user-info model: Keep 2 input variables independent to avoid affecting each other. See figure.

The Method - Model Architecture

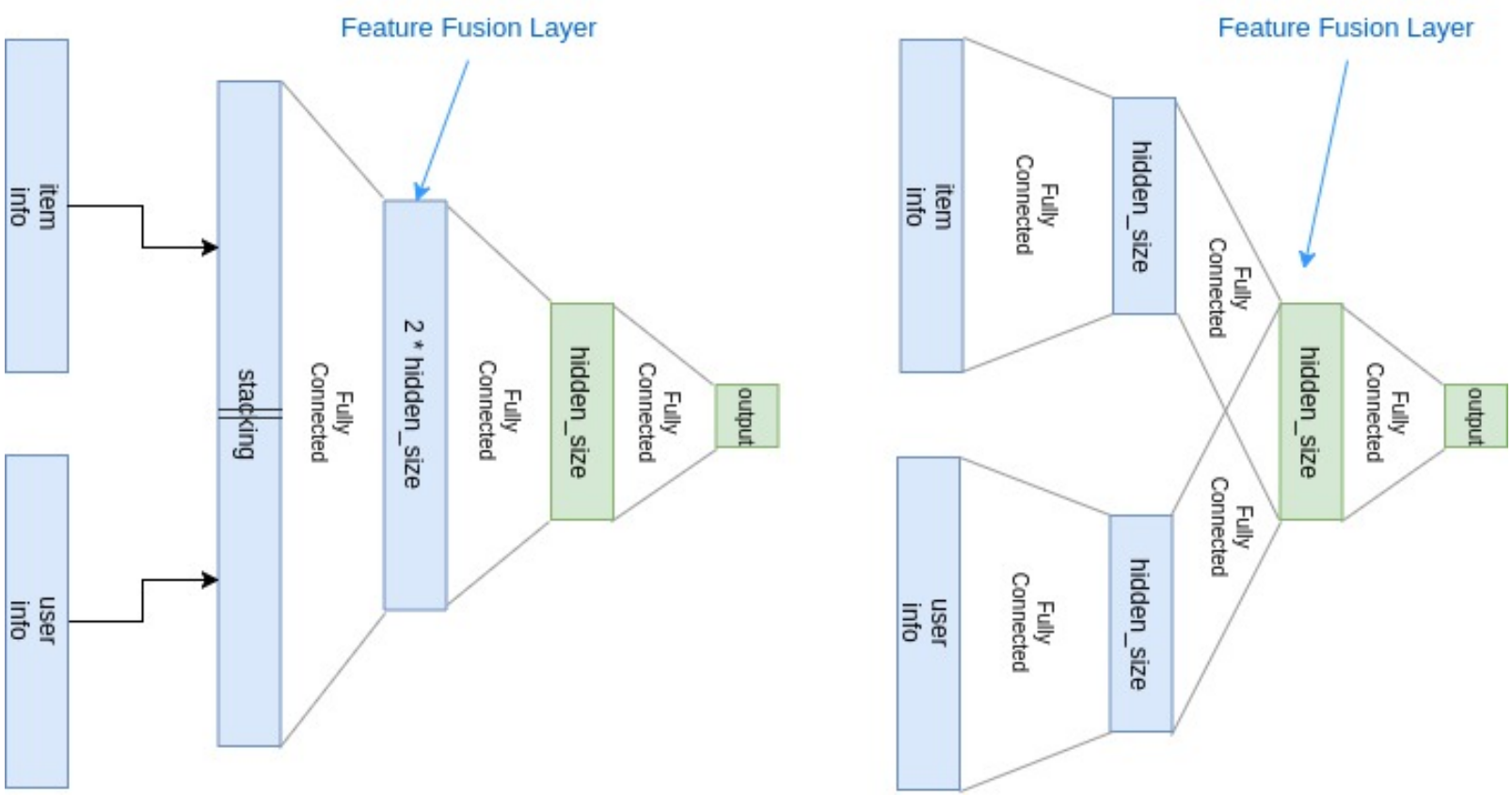
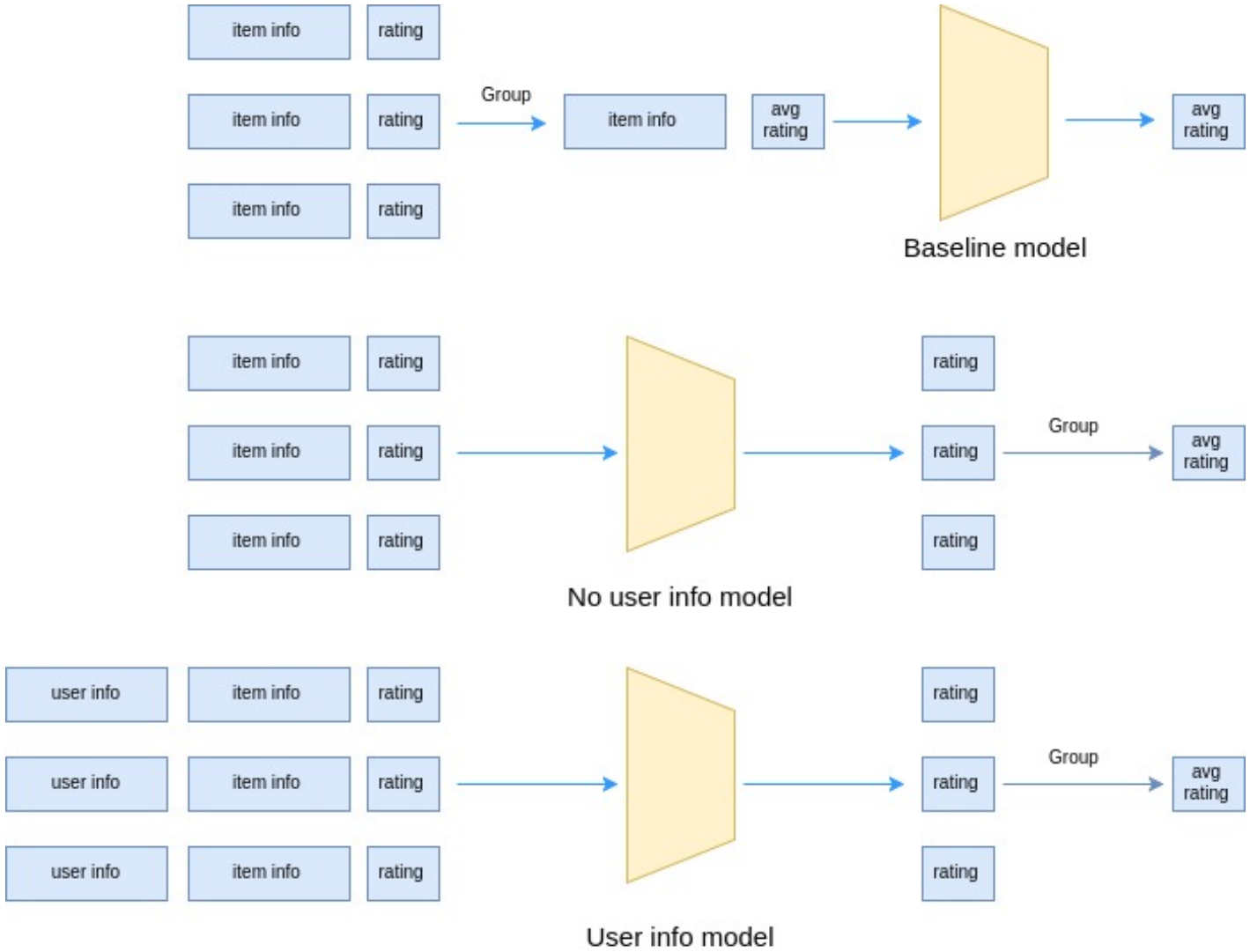


Figure: Architecture of Stacking User-info Model (left) and Embedding User-info Model (Right). The blue block represents normal vector, while the green indicates applying Batch Normalization before activation function.

Experiment and Results

1. Baseline vs No-user info model
2. No user-info model vs User info models
3. Stacking user-info vs Embedding user-info

Experiment and Results



Experiment and Results

1. Baseline
   - Training dataset: Grouping ratings by item_id
   - Objective: Mean ratings of each item.
   - Evaluate: RMSE.
2. No user-info model
   - Training dataset: All ratings with only item_info.
   - Objective: Exact value of each rating.
   - Evaluate: grouping by item_id, get mean ratings of each item, calculate RMSE.
3. Stacking user-info model, Embedding user-info model
   - Training dataset: All ratings with item_info and user_info
   - Objective: Exact value of each rating.
   - Evaluate: grouping by item_id, get mean ratings of each item, calculate RMSE.

Experiment and Results

Model	RMSE
Baseline	0.6170
No user-info	0.5986
Stacking user-info	0.5890
Embedding user-info	0.5866

Table: The result of model

Experiment and Results

	Reference model	Relative Change
No user-info	Baseline	-2.99%
Stacking user-info	Baseline	-4.54%
Embedding user-info	Baseline	-4.93%
Stacking user-info	No user-info	-1.59%
Embedding user-info	No user-info	-2.00%
Embedding user-info	Stacking user-info	-0.41%

Table: The result of model

Summary

- Predict exact values rather than mean of ratings
  Nearly 3% error decreases.
- Adding user information
  Improve 2% error.
- Using ANN
  Many flexible ways to add user information.
  Embedding is better than stacking.