Collaborative Filtering Applied to Educational Data Mining

Andreas Töscher, Michael Jahrer
commendo research & consulting GmbH, AUSTRIA

Abstract

We present our overall third ranking solution for the KDD Cup 2010 on educational data mining. The goal of the competition was to predict a student's ability to answer questions correctly, based on historic results. In our approach we use an ensemble of collaborative filtering techniques, as used in the field of recommender systems and adopt them to fit the needs of the competition. The ensemble of predictions is finally blended, using a neural network.

K Nearest Neighbor (KNN)

$$\bar{\rho}_{s_1 s_2} = \frac{|\mathbb{I}_{s_1 s_2}| \cdot \rho_{s_1 s_2}}{|\mathbb{I}_{s_1 s_2}| + \alpha}$$

$$\sigma(x) = \frac{1}{1 - e^{-x}}$$

$$\tilde{\rho}_{s_1 s_2} = \sigma\left(\delta \cdot \bar{\rho}_{s_1 s_2} + \gamma\right)$$

$$\tilde{c}_{is} = \frac{\sum_{\tilde{s} \in \mathbb{S}_i(s;K)} \tilde{\rho}_{s\tilde{s}} c_{i\tilde{s}}}{\sum_{\tilde{s} \in \mathbb{S}_i(s;K)} |\tilde{\rho}_{s\tilde{s}}|}$$

$$\hat{c}_{is} = \frac{\tilde{c}_{is} \sum_{\tilde{s} \in \mathbb{S}_i(s;K)} |\tilde{\rho}_{s\tilde{s}}| + \mu_s \beta}{\sum_{\tilde{s} \in \mathbb{S}_i(s;K)} |\tilde{\rho}_{s\tilde{s}}| + \beta}$$

- $\mathbb{I}_{s_1 s_2}$ ... set of steps commonly answered by student s_1 and student s_2
- $\hat{c}_{is}$ ... prediction for student s on step i
- $\rho_{s_1 s_2}$ ... Pearson correlation between student s_1 and student s_2
- $\mathbb{S}_i(s;K)$ ... to student s the K most similar students who answered step i
- $\alpha, \beta, \gamma, \delta$ ... meta parameters

Dataset	RMSE	Meta parameters
Algebra 2008-2009	0.3257	$K = 41, \alpha = 12.9, \beta = 1.5, \delta = 6.2, \gamma = -1.9$
Bridge to Algebra 2008-2009	0.3049	$K = 41, \alpha = 12.9, \beta = 1.5, \delta = 6.2, \gamma = -1.9$

Factor Model 1 (FM1)

$$\hat{c}_{is} = \mu + \hat{\mu}_i + \bar{\mu}_s + \tilde{\mu}_{p(i)} + \acute{\mu}_{x(i)} + \breve{\mu}_{u(i)} + \frac{1}{\sqrt{|\mathbb{K}(i,s)|}} \sum_{k \in \mathbb{K}(i,s)} (\tilde{\mu}_k + \acute{\mu}_{ks}) + \left(\mathbf{a}_i + \frac{1}{\sqrt{|\mathbb{K}(i,s)|}} \sum_{k \in \mathbb{K}(i,s)} \acute{\mathbf{a}}_k\right)^T \cdot \mathbf{b}_s$$

- μ ... global bias
- $\hat{\mu}_i$ ... bias for step i
- $\bar{\mu}_s$ ... bias for student s
- $\tilde{\mu}_{p(i)}$ ... bias for problem $p(i)$
- $\acute{\mu}_{x(i)}$ ... bias for section $x(i)$
- $\breve{\mu}_{u(i)}$ ... bias for unit $u(i)$
- $\mathbb{K}(i,s)$ ... set of knowledge components
- $\acute{\mathbf{a}}_k$ ... N dimensional KC feature vector

Dataset	RMSE	Meta parameters
Algebra 2008-2009	0.3078	$N = 50, \eta = 0.0005, \lambda = 0.01$
Bridge to Algebra 2008-2009	0.3013	$N = 50, \eta = 0.0005, \lambda = 0.01$

Factor Model 2 (FM2)

$$\hat{c}_{is} = \mu + \hat{\mu}_i + \bar{\mu}_s + \tilde{\mu}_{p(i)} + \breve{\mu}_{u(i)} + \ddot{\mu}_{su(i)} + \frac{1}{\sqrt{|\mathbb{K}(i,s)|}} \sum_{k \in \mathbb{K}(i,s)} \tilde{\mu}_k + \left(\mathbf{a}_i + \frac{1}{\sqrt{|\mathbb{K}(i,s)|}} \sum_{k \in \mathbb{K}(i,s)} \acute{\mathbf{a}}_k\right)^T \cdot \mathbf{b}_s$$

- $\hat{\mu}_i$ ... bias for step i
- $\bar{\mu}_s$ ... bias for student s
- $\tilde{\mu}_{p(i)}$ ... bias for problem $p(i)$
- $\acute{\mu}_{x(i)}$ ... bias for section $x(i)$
- $\breve{\mu}_{u(i)}$ ... bias for unit $u(i)$
- $\ddot{\mu}_{su(i)}$ ... student s dependent unit bias
- $\mathbb{K}(i,s)$ ... set of knowledge components
- $\acute{\mathbf{a}}_k$ ... N dimensional KC feature vector

Dataset	RMSE	Meta parameters
Algebra 2008-2009	0.3044	$N = 150, \eta = 0.002, \lambda = 0.005$
Algebra 2008-2009	0.3045	$N = 150, \eta = 0.003, \lambda = 0.003$
Bridge to Algebra 2008-2009	0.2997	$N = 100, \eta = 0.003, \lambda = 0.005$

Group Factor Model (GFM)

$$\hat{c}_{is} = \mu + \sum_{g=1}^{7} \mathbf{b}_s^T \cdot \mathbf{d}_{g,G^{(g)}(i,s)} + \sum_{g=1}^{7} \bar{\mathbf{a}}_i^T \cdot \bar{\mathbf{d}}_{g,G^{(g)}(i,s)} + \sum_{g=1}^{7} \tilde{\mathbf{b}}_{sg}^T \cdot \tilde{\mathbf{d}}_{g,G^{(g)}(i,s)} + \sum_{g=1}^{7} \hat{\mathbf{a}}_{ig}^T \cdot \hat{\mathbf{d}}_{g,G^{(g)}(i,s)} + \sum_{g=1}^{7} \sum_{\hat{g}=1}^{7} \acute{\mathbf{d}}_{g,G^{(g)}(i,s)}^T \cdot \acute{\mathbf{d}}_{\hat{g},G^{(\hat{g})}(i,s)} + \sum_{g=1}^{7} \sum_{\hat{g}=g+1}^{7} \check{\mathbf{d}}_{g,G^{(g)}(i,s)}^T \cdot \check{\mathbf{d}}_{\hat{g},G^{(\hat{g})}(i,s)}$$

- 7 groups: unit, section, problem name, plain step name, knowledge component, opportunity count, problem view
- $G^{(g)}(i,s)$ ... group g dependent mapping function, which maps a given student s step i pair to a group ID
- $\tilde{\mathbf{b}}_{sg}$ ... N dimensional feature vector for student s and group g
- $\hat{\mathbf{a}}_{ig}$ ... N dimensional feature vector for step i and group g
- $\mathbf{d}_{g,G^{(g)}(i,s)}$ ... N dimensional feature vector for group g and group ID $G^{(g)}(i,s)$

Dataset	RMSE	Meta parameters
Algebra 2008-2009	0.2997	$N = 20, \eta = 0.003, \eta^- = 0.0001, \lambda = 0.001$
Algebra 2008-2009	0.3051	$N = 20, \eta = 0.002, \eta^- = 0.0001, \lambda = 0.0$
Bridge to Algebra 2008-2009	0.2908	$N = 20, \eta = 0.005, \eta^- = 0.0002, \lambda = 0.0016$
Bridge to Algebra 2008-2009	0.2901	$N = 50, \eta = 0.005, \eta^- = 0.0002, \lambda = 0.0016$
Bridge to Algebra 2008-2009	0.2898	$N = 60, \eta = 0.005, \eta^- = 0.0002, \lambda = 0.0016$

Singular Value Decomposition (SVD)

$$\hat{c}_{is} = \mathbf{a}_i^T \cdot \mathbf{b}_s$$

$$E = \sum_{(i,s) \in \mathbb{L}_T} (c_{is} - \hat{c}_{is})^2 + \lambda(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2)$$

- $\mathbf{a}_i$ ... N dimensional feature vector for step i
- $\mathbf{b}_s$ ... N dimensional feature vector for student s
- $\hat{c}_{is}$ ... prediction for student s on step i
- λ ... meta parameter controlling the L2 regularization
- $\mathbb{L}_T$ ... set of training examples

Dataset	RMSE	Meta parameters
Algebra 2008-2009	0.446277	$N = 10, \eta = 0.002, \lambda = 0.02$
Bridge to Algebra 2008-2009	0.3168	$N = 10, \eta = 0.002, \lambda = 0.02$
Bridge to Algebra 2008-2009	0.3159	$N = 20, \eta = 0.002, \lambda = 0.01$
Bridge to Algebra 2008-2009	0.3178	$N = 20, \eta = 0.002, \lambda = 0.03$

Factor Model 3 (FM3)

$$\breve{\mu}_{is} = \mu + \hat{\mu}_i + \bar{\mu}_s + \tilde{\mu}_{p(i)} + \acute{\mu}_{x(i)} + \breve{\mu}_{u(i)} + \frac{1}{\sqrt{|\mathbb{K}(i,s)|}} \sum_{k \in \mathbb{K}(i,s)} \breve{\mu}_k$$

$$\breve{\mathbf{a}}_{is} = \mathbf{a}_i + \hat{\mathbf{a}}_{u(i)} + \bar{\mathbf{a}}_{x(i)} + \tilde{\mathbf{a}}_{p(i)} + \frac{1}{\sqrt{|\mathbb{K}(i,s)|}} \sum_{k \in \mathbb{K}(i,s)} \acute{\mathbf{a}}_k$$

$$\breve{\mathbf{b}}_{is} = \mathbf{b}_s + \hat{\mathbf{b}}_{u(i)} + \bar{\mathbf{b}}_{x(i)} + \tilde{\mathbf{b}}_{p(i)} + \frac{1}{\sqrt{|\mathbb{K}(i,s)|}} \sum_{k \in \mathbb{K}(i,s)} \acute{\mathbf{b}}_k$$

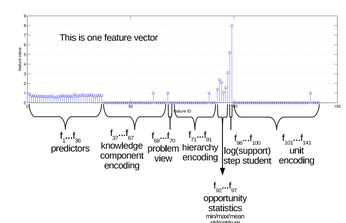
$$\hat{c}_{is} = \breve{\mu}_{is} + \breve{\mathbf{a}}_{is}^T \cdot \breve{\mathbf{b}}_{is}$$

- $\mathbb{K}(i,s)$ ... set of knowledge components
- $\breve{\mu}_k$ ... KC bias
- $\mathbf{a}_i$ ... N dimensional feature vector for step i
- $\mathbf{b}_s$ ... N dimensional feature vector for student s
- $\hat{\mathbf{a}}_{u(i)}, \hat{\mathbf{b}}_{u(i)}$ ... feature vectors for unit $u(i)$
- $\bar{\mathbf{a}}_{x(i)}, \bar{\mathbf{b}}_{x(i)}$ ... feature vectors for section $x(i)$
- $\tilde{\mathbf{a}}_{p(i)}, \tilde{\mathbf{b}}_{p(i)}$ ... feature vectors for problem $p(i)$
- $\acute{\mathbf{a}}_k, \acute{\mathbf{b}}_k$ ... KC feature vectors

Dataset	RMSE	Meta parameters
Algebra 2008-2009	0.2996	$N = 5$
Bridge to Algebra 2008-2009	0.2916	$N = 20$
Bridge to Algebra 2008-2009	0.2924	$N = 5$

Blending

The training set for the blender consists of the predictions we get from the cross validation process from each individual collaborative filtering algorithm. For the first dataset (Algebra 2008-2009) we use 36 predictors, for the second dataset (Bridge to Algebra 2008-2009) 37 predictors are used. After a few experiments with different learners and different parameters it turned out that a neural network with two hidden layer works best. The net layout for the first dataset is 142-80-80-1 and for the second dataset 130-80-80-1. Both blending networks are trained for 90 epochs, the learning rate $\eta = 0.001$ is subtracted by 10^{-5} every epoch. No weight decay is used. The features in the training set, which are now predictions from CF algorithms, are enlarged by additional information. We add knowledge component encoding, problem view encoding, hierarchy encoding, unit encoding, opportunity statistics and the student/step support. Encoding is a binary information source, where exactly one input has value 1.0.



	Dataset: Algebra 2008-2009	Dataset: Bridge to Algebra 2008-2009
blend	0.280925: on cross-validation	0.288417: on cross-validation
blend	0.277355: submission	0.28073: submission