从 KDDCUP2012 看微博好友推荐

Chunwei Yan

superjom@sz.pku.edu.cn

June 3, 2013

背景

- KDDCUP2012
- 腾讯微博
- 1st-0.4265 undergrads@ 数据和知识管理实验室 @ 上海交大
- 2st-0.41874 盛大研究院

Outline

- 训练数据格式
- 超越矩阵分解模型
- 其他方法
- 实验分析

KDDCUP 2012 Track1

- 目标: 用户好友推荐, MAP@3

KDDCUP 2012 Track1

- 目标: 用户好友推荐, MAP@3
- datasets:
  1. 训练集:(UserId)(ItemId)(Result)(Unix-timestamp)
  2. 其他信息:
     - Profile: (UserId)(Year-of-birth)(Gender)(Number-of-tweet)(Tag-Ids)
     - item: (ItemId)(Item-Category)(Item-Keyword)
     - user-action: (UserId)(Action-Destination-UserId)(Number-of-at-action)(Number-of-retweet)(Number-of-comment)

矩阵分解模型 (SVD/SVD++)

$$\hat{r}_{ui} = (\sum_{c \in C(u)} \alpha_c^{(u)} \mathbf{p}_c)^T (\sum_{c \in C(i)} \beta_c^{(i)} \mathbf{q}_c) + \sum_{c \in C(u,i)} \gamma_c^{(u,i)} g_c \tag{1}$$

矩阵分解模型 (SVD/SVD++)

$$\hat{r}_{ui} = \Big(\sum_{c \in C(u)} \alpha_c^{(u)} \mathbf{p}_c\Big)^T \Big(\sum_{c \in C(i)} \beta_c^{(i)} \mathbf{q}_c\Big) + \sum_{c \in C(u,i)} \gamma_c^{(u,i)} g_c \qquad (1)$$

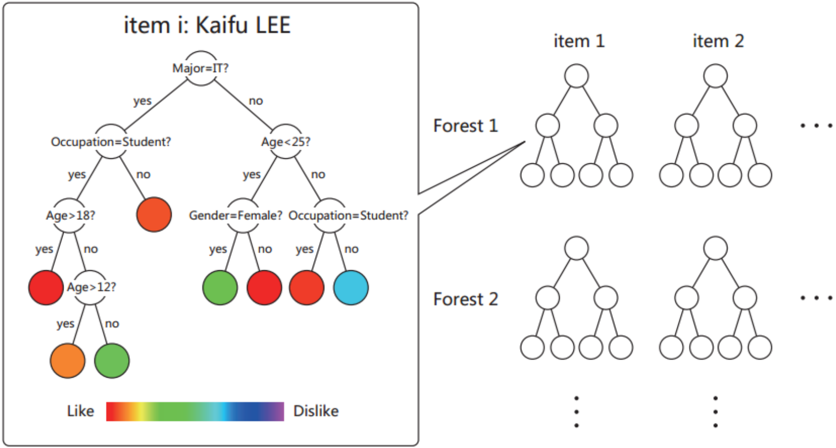
- $\theta = \{\mathbf{p}, \mathbf{q}, g\}$. 模型参数, 随机梯度下降
- $\alpha_c^{(u)}$ 用户特征 (user features), tags, keywords, 社交网络
- $\beta_c^{(i)}$ item features, 分类, 网络
- $\gamma_c^{(u,i)}$ 公共特征, user, item 间的交互

Additive Forest

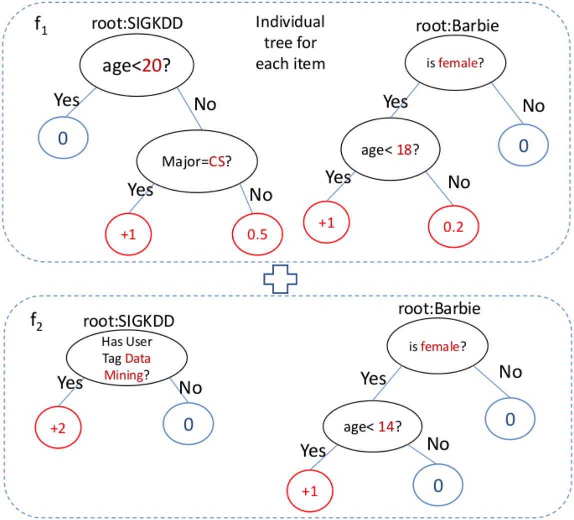
$$\hat{r}_{ui} = \sum_{s=1}^{S} f_{s,root(i,s)}(x_u) \tag{2}$$

- x_u 用户 u 的特性
- $f_{s,root(i,s)}$ 用回归树定义的函数
- 采用梯度提升法学习

Additive Forest



Additive Forest 实例



矩阵分解模型 vs Additive Forest

	矩阵分解	Additive Forest
稀疏矩阵处理	非常好	一般
不同信息整合	线性组合	非线性组合
对连续值的处理	人为划分	自动产生划分

矩阵分解模型 vs Additive Forest

	矩阵分解	Additive Forest
稀疏矩阵处理	非常好	一般
不同信息整合	线性组合	非线性组合
对连续值的处理	人为划分	自动产生划分

- 两种模型都各有各自的特点
- 结合他们的特点对提高精度至关重要

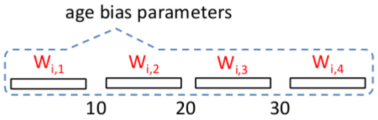
矩阵分解模型 vs Additive Forest

矩阵分解模型 vs Additive Forest

矩阵分解模型

$$\hat{r}_{ui} = p_u^T q_i + W_{i,ag(u)} \qquad (3)$$

- $ag(u)$ 年龄划分索引
- 需要实现人为划分

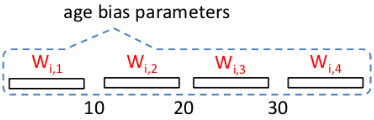


矩阵分解模型 vs Additive Forest

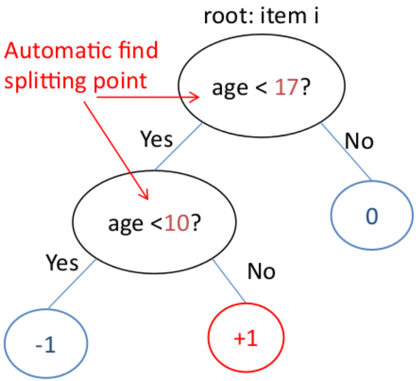
矩阵分解模型

$$\hat{r}_{ui} = p_u^T q_i + W_{i,ag(u)} \qquad (3)$$

- $ag(u)$ 年龄划分索引
- 需要实现人为划分



Additive Forest



社交网络

$$\hat{r}_{ui} = \left(\frac{1}{\sqrt{|F(u)|}} \sum_{j \in F(u)} \mathbf{p}_j\right)^T \mathbf{q}_i + b_i \tag{4}$$

社交网络

$$\hat{r}_{ui} = \left(\frac{1}{\sqrt{|F(u)|}} \sum_{j \in F(u)} \mathbf{p}_j\right)^T \mathbf{q}_i + b_i \tag{4}$$

- $F(u)$ user u follow 的好友
- 模拟其好友对其影响

关键词和 Tag

- 分类信息对用户的预测有影响
- 影响无法无法计量

关键词和 Tag

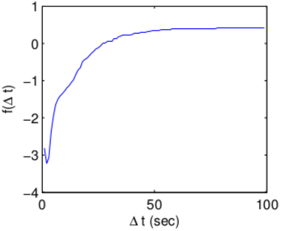
- 分类信息对用户的预测有影响
- 影响无法无法计量
- 作为潜在因素 (SVD++)

$$\mathbf{p}'_u = \mathbf{p}_u + \frac{1}{||w_u||_2} \sum_{j \in K(u)} w_{u,j} \mathbf{y}_j \tag{5}$$

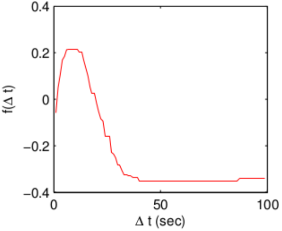
- $K(u)$ 用户 u 的 keywords or tags
- $w_{u,j}$ 特征的权重

用户序列模式 (User Sequential Patterns)

- 推测用户点击趋势



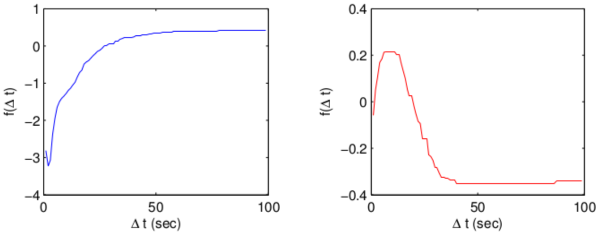
(a) $\Delta t = t_{next} - t_{curr}$



(b) $\Delta t = t_{curr} - t_{prev}$

用户序列模式 (User Sequential Patterns)

- 推测用户点击趋势



(a) $\Delta t = t_{next} - t_{curr}$ (b) $\Delta t = t_{curr} - t_{prev}$

$$\hat{r}'_{ui}(t) = \hat{r}_{ui} + f(\triangle t), f(\triangle t) = \sum_{s=1}^{S} f_s(\triangle t) \tag{6}$$

实验结果

ID	model	public	private	Δ_{public}	$\Delta_{private}$
1	item bias	34.6%	34.0%		
2	1 + user follow/action	36.7%	35.8%	2.1%	1.8%
3	2 + user age/gender	38.0%	37.2%	1.3%	1.4%
4	3 + user tag/keyword	38.5%	37.6%	0.5%	0.4%
5	4 + item taxonomy	38.7%	37.8%	0.2%	0.2%
6	5 + time-aware model	39.0%	37.9%	0.3%	0.1%
7	6 + age/gender(forest)	39.1%	38.0%	0.1%	0.1%
8	7 + sequential patterns	44.2%	42.7%	5.1%	4.7%

Table: MAP@3 of different methods

引用

[Tianqi Chen, Linpeng Tan, QIn Liu and so on,2012]
ACM
Combining Factorization Model and Additive Forest for Collaborative Followee Recommendation, 2004

[Yehuda Koren, 2008]
Factorization Meets the Neighborhood: a Multifaceted Collborative Filtering Model, 2008

Thank You, Questions?