Построение ансамбля алгоритмов рекомендаций

Выпускная квалификационная работа

Выполнил:
студент 417 группы
Кудрявцев Георгий Алексеевич

Научный руководитель:
д.ф-м.н., профессор
Дьяконов Александр Геннадьевич

3 мая 2016 г.

Неформальная постановка задачи

Требуется улучшить качество работы алгоритмов ранжирования при помощи ансамблирования уже существующих методов.

Предметная область

Рассматривается задачи ранжирования по данным с двоичной релевантностью.
На практике данная задача решается при помощи алгоритмов машинного обучения.
В данной работе рассматриваются факторизационные методы и их линейные ансамбли.

Актуальность задачи

Построение рекоммендаций для:

- социальных сетей
- сайтов знакомств
- интернет магазинов

Цель и задачи

Цель: разработать метод ансамблирования, который стабильно улучшает качество ранжирования.
Задачи:

- Составить обзор современных факторизационных методов ранжирования.
- Предложить эффективный метод ансамблирования.
- Реализовать методы и провести их сравнительный анализ.

Формальная постановка задачи

Входные данные:
матрица R размера $M \times N$, где M - количество пользователей, N - количество предметов. $R_{ui} = 1$, если пользователей u взаимодействовал с предметов i. В противном случае $R_{ui} = 0$.
Выходные данные:
Для каждого пользователя u ранжированный список предметов, которые не лежат в тренировочной выборке.

Обзор существующих методов

- **CLiMF**[1] – Факторизационный метод, который оптимизирует сглаженную версию метрики MRR.
- **MPR_MF**[2] – Факторизационный метод, который оптимизирует AUC.
- **TFMAP**[3] – Факторизационный метод, который оптимизирует сглаженную метрику MAP.
- **iMF**[4] – Факторизационный метод, который оптимизирует взвешенную квадратичную ошибку.

---

[1] Yue Shi, Alexandros Karatzoglou, Linas Baltrunas. CLiMF: learning to maximize reciprocal rank with collaborative less-is-more filtering. 2012

[2] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner. BPR: Bayesian Personalized Ranking from Implicit Feedback. 2009

[3] Yue Shia,Alexandros Karatzogloub, Linas Baltrunas. TFMAP: Optimizing MAP for Top-N Context-aware Recommendation. 2012

[4] Yifan Hu, Yehuda Koren, Chris Volinsky. Collaborative Filtering for Implicit Feedback Datasets. 2008

Линейный ансамбль

Пусть имеется множество базовых алгоритмов $b_i(x)$.
Необходимо подобрать такие веса α_i, чтобы линейная комбинация алгоритмов $\hat{b}(x) = \sum_i \alpha_i b_i(x)$ показывала лучший результат по какому-нибудь заданному функционалу.

Оптимальное взвешенное голосование

Рассмотрим линейную комбинацию двух алгоритмов ранжирования в следующем виде:

$$\hat{f}_{ui} = \alpha f_{ui}^{m_1} + (1-\alpha) f_{ui}^{m_2} \text{ , где } 0 \leq \alpha \leq 1$$

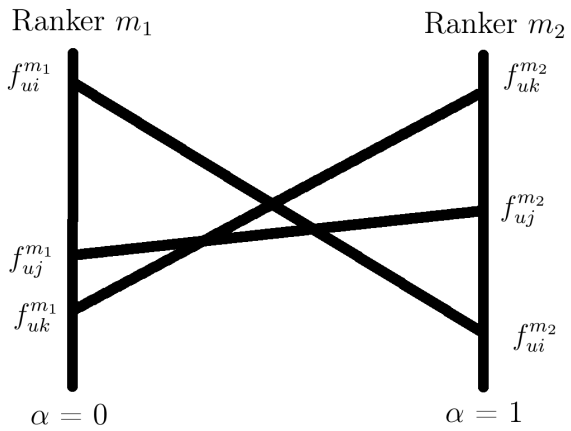

Рис. : Графическая демонстрация идеи метода [5]

[5] Qiang Wu, Christopher J. C. Burges, Krysta M. Svore. Adapting Boosting for Information Retrieval Measures. 2010.

Экспериментальные оценка

MovieLens 100k										
	ptest = 10%, pvalid=10%, maxiter = 5					ptest = 20%, valid = 10%, maxiter = 5				
	BSM	SV	RV	OVB	OVT	BSM	SV	RV	OVB	OVT
P@5	0.2965	0.1602	0.2528	**0.3069**	0.3019	0.4451	0.2531	0.3831	**0.4498**	0.4464
1call@5	0.7344	0.4950	0.6712	**0.7580**	0.7555	0.8706	0.6203	0.8039	**0.8784**	0.8734
NDCG@5	0.3201	0.1778	0.2783	**0.3307**	0.3257	0.4705	0.2738	0.4070	**0.4770**	0.4752
MAP@5	0.5035	0.3239	0.4599	**0.5185**	0.5152	0.6480	0.4261	0.5763	0.6555	**0.6562**
Epinion										
	ptest = 10%, pvalid=10%, maxiter = 5					ptest = 20%, valid = 10%, maxiter = 5				
	BSM	SV	RV	OVB	OVT	BSM	SV	RV	OVB	OVT
P@5	0.0716	0.0295	0.0534	**0.0764**	0.0760	0.1283	0.0550	0.1000	0.1328	**0.1363**
1call@5	0.2710	0.1303	0.2183	**0.2921**	0.2908	0.4179	0.2242	0.3595	0.4372	**0.4438**
NDCG@5	0.0775	0.0334	0.0572	**0.0830**	0.0822	0.1384	0.0615	0.1062	0.1425	**0.1468**
MAP@5	0.1561	0.0759	0.1193	**0.1688**	0.1672	0.2536	0.1316	0.2035	0.2607	**0.2686**

Экспериментальные оценка

Slashdot										
	ptest = 10%, pvalid=10%, maxiter = 5					ptest = 20%, valid = 10%, maxiter = 5				
	BSM	SV	RV	OVB	OVT	BSM	SV	RV	OVB	OVT
P@5	0.0431	0.0186	0.0291	**0.0438**	0.0436	0.0746	0.0362	0.0519	0.0757	**0.0758**
1call@5	0.1644	0.0866	0.1272	**0.1683**	0.1671	0.2570	0.1577	0.2038	0.2631	**0.2653**
NDCG@5	0.0473	0.0201	0.0312	**0.0481**	**0.0481**	0.0803	0.0397	0.0556	0.0819	**0.0822**
MAP@5	0.0959	0.0455	0.0684	0.0976	**0.0978**	0.1506	0.0877	0.1140	0.1550	**0.1562**
Movie Lens 1m										
	ptest = 10%, pvalid=10%, maxiter = 5					ptest = 20%, valid = 10%, maxiter = 5				
	BSM	SV	RV	OVB	OVT	BSM	SV	RV	OVB	OVT
P@5	0.2601	0.1065	0.2374	0.2628	**0.2639**	0.3936	0.1974	0.3616	**0.3983**	0.3944
1call@5	0.6541	0.3424	0.6186	0.6566	**0.6593**	0.7985	0.5255	0.7714	**0.8075**	0.8066
NDCG@5	0.2795	0.1104	0.2476	0.2830	**0.2836**	0.4127	0.2007	0.3710	**0.4192**	0.4157
MAP@5	0.4384	0.1936	0.3859	0.4440	**0.4438**	0.5729	0.3127	0.5166	**0.5839**	0.5822

Результаты работы

1. Составлен обзор основных факторизационных методов для задачи ранжирования для набора данных с двоичной релевантностью.
2. Предложен метод ансамблирования, который стабильной улучшает качество работы ранжирования.
3. Реализованы факторизационные методы и эксперименты на языке python и c++.