h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Session-Based-Recommendersysteme für Möbelprodukte

Sebastian Reimer

Hochschule Darmstadt, Fachbereiche Mathematik und Naturwissenschaften & Informatik

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Motivation

Artikelempfehlungen des kooperierenden Unternehmens für diese Masterarbeit werden aktuell eimal wöchentlich pro Artikel berechnet und auf der Artikelseite ausgespielt. Gelangen User:innen auf eine Artikelseite, so sind diese Empfehlungen für alle User:innen identisch, unabhängig vom bisherigen Session-Verlauf der User:innen. Daher wurden elf Vergleichsmodelle getestet, die in der Lage sind auf Basis des bisherigen Session-Verlaufs eines/einer User:in Empfehlungen zu erstellen. Die Ziele dieser Arbeit waren:

1. Es soll mindestens eines der elf Vergleichsmodelle bessere Ergebnisse in Bezug auf die Metriken MRR@20 und HR@20 erzielen. Wünschenswert sind ebenfalls bessere Resultate für die Metriken Coverage@20 und Popularity@20.
2. Um zusätzlich den Einfluss von Sessionlängen bewerten zu können, sollen die Modelle mit drei Testdatensätzen nach den oben genannten Kriterien evaluiert werden: Der Erste enthält alle Sessionlängen, der Zweite enthält lange Sessions und der Dritte enthält kurze Sessions.
3. Des Weiteren soll geprüft werden, wie häufig die verwendeten Modelle Fehlklassifizierungen für die Produktkategorien machen. Auf diese Weise kann evaluiert werden, welches Modell sich für welche Produktkategorie am Besten eignet.
4. Aufdecken von Mustern zur Entscheidungsfindung zwischen dem besten Vergleichsmodell sowie dem bestehenden Modell mithilfe von Vorhersagedifferenzen.

Methoden

Modelle

Es wurden für den Vergleich mit dem bestehenden Modell (Mahout) elf Vergleichsmodelle verwendet:

- Popularity-Modell (POP)
- Session-Popularity-Modell (S-POP)
- Sequential Rules (SR)
- Item-basierter-k-Nächste-Nachbarn Algorithmus (IKNN)
- Session-basierter-k-Nächste-Nachbarn Algorithmus (SKNN)
- Sequence and Time Aware Neighborhood (STAN)
- Vector STAN (VSTAN)
- Vector Multiplication Session-Based k-Nächste-Nachbarn (V-SKNN)
- Gated Recurrent Units for Recommendations (GRU4REC)
- Session-Based-Recommendersystem mit Graph Neural Networks (SRGNN)
- Efficient Manifold Density Estimator (EMDE)

Berechnung der Vorhersagedifferenzen

Die Berechnung der Vorhersagedifferenzen erfolgte auf Basis des Testdatensatzes mit allen Sessionlängen. Es wurden nur die Vorhersagevorgänge ausgewählt, bei denen sowohl Mahout als auch das jeweilige Vergleichsmodell pro Vorhersagevorgang für ein Item eine Gewichtung berechnet hatte. Wie der Abbildung 1 zu entnehmen ist, erfolgte nach der Filterung eine Kalibierung pro Vorhersagevorgang in den Wertebereich $[0, 1]$. Schließlich konnte die Berechnung der Vorhersagedifferenz Δ des Items k von der Gewichtung w_M des Mahout-Modells sowie der Gewichtung w_V des Vergleichsmodells pro Vorhersagevorang wie folgt berechnet werden: $\Delta = w_M(k) - w_V(k)$.

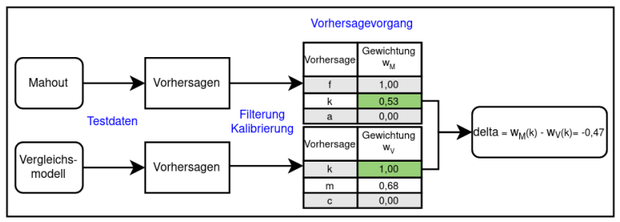

Abbildung 1. Vorgehensweise zur Berechnung der Vorhersagedifferenzen.

In einer Analyse wurden für das SR-GNN-Modell pro Position des vorherzusagenden Items die durchschnittlichen Δ-Werte für die jeweilige Positionen des vorherzusagenden Items in der jeweiligen Session aus dem Testdatensatz berechnet. Dies wurde für die falschen sowie für die korrekt vorhergesagten Items durchgeführt.

Ergebnisse

Resultate des kompleten Testdatensatz

Für die Ergebnisse in Tabelle 1 wurden alle Sessionlängen verwendet. Aus dieser Tabelle kann entnommen werden, dass SR-GNN die besten Werte in Bezug auf MRR@20 und HR@20 lieferte. Bei COV@20 und Pop@20 schnitt Mahout am zweitbesten bzw. besten ab.

Tabelle 1. Ergebnisse des Durchlaufs mit allen Sessions aus dem Testdatensatz. Fettgedruckt sind pro Metrik jeweils die **besten** Werte. Unterstrichen sind pro Metrik jeweils die zweitbesten Werte.

Modell	MRR@20	HR@20	Cov@20	Pop@20	Trainingsdauer
Mahout	0,2390	0,5009	0,9774	**0,0161**	58min 18s
IKNN	0,2402	0,5037	**0,9797**	0,0195	4min 29s
POP	0,0085	0,0298	0,0003	0,3563	0,32s
S-POP	0,2793	0,4402	0,7935	0,2998	0,28s
SR	0,3825	0,7142	0,9612	0,0362	39s
SKNN	0,3174	0,6332	0,8930	0,0435	14s
V-SKNN	0,3264	0,6282	0,9262	0,0487	13s
STAN	0,3546	0,6816	0,9299	0,0371	15s
VSTAN	0,3496	0,6801	0,9504	0,0282	16s
GRU4REC	0,3463	0,7095	0,9506	0,0165	4h 6min 29s
SR-GNN	**0,4045**	**0,7231**	0,6995	0,0373	10h 2min 55s
EMDE	0,3851	0,7009	0,9526	0,0234	1h 20min 33s

Aus der vorherigen Tabelle wird ersichtlich, dass SR-GNN in Bezug auf die Genauigkeitsmetriken MRR@20 und HR@20 die besten Werte lieferte. Dies war ebenfalls der Fall die Evaluierung mit kurzen sowie mit langen Sessions (Ziel 2). Darüber hinaus überzeugte SR-GNN als bestes Modell in Bezug auf die Falschklassifizierungen für die Produktkategorien (Ziel 3).

Vorhersagedifferenzen für korrekte Vorhersagen

In Abbildung 1 sind die Vorhersagedifferenzen für SR-GNN für alle korrekt klassifizierten Vorhersagevorgänge in einem 2-dimensionalen Histogramm zusammen mit den relativen Häufigkeiten des vorherzusagenden Items im Trainingsdatensatz dargestellt. Im Bereich von $\Delta = 0$ befanden sich die meisten korrekten Vorhersagevorgänge, wobei ebenfalls ein schwächer werdender Farbverlauf in Richtung -1 und $+1$ zu erkennen war, der für SR-GNN etwas stärker ausfällt, es also mehr korrekte Vorhersagevorgänge gibt. Besonders auffällig ist, dass sich im oberen rechten Bereich keine Datenpunkte befanden. Lediglich im negativen Vorhersagedifferenzbereich befanden sich für häufig vorkommende Items im Trainingsdatensatz korrekte Vorhersagen.

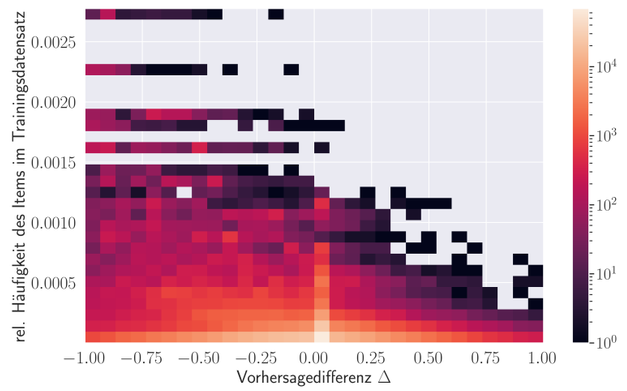

Abbildung 2. 2-dimensionales Histogramm mit der relativen Häufigkeit des vorherzusagenden Items im Trainingsdatensatz sowie der Vorhersagedifferenz Δ des SR-GNN-Modells für die korrekt vorhergesagten Items.

Vorhersagedifferenzen mit Position innerhalb einer Session

Der Abbildung 3 kann entnommen werden, dass die durchschnittlichen Vorhersagedifferenzen für die korrekt vorhergesagten Items bei Position 1 bis zu ca. 30 im negativen Bereich sowie unterhalb der Kurve der falschen Vorhersagevorgänge lagen. Darüber hinaus kann eine Erhöhung der durchschnittlichen Vorhersagedifferenzen bei größer werdender Position für die korrekt klassifizierten Items erkannt werden. Bei den falsch vorhergesagten Items war die Kurve nahezu konstant parallel zur X-Achse.

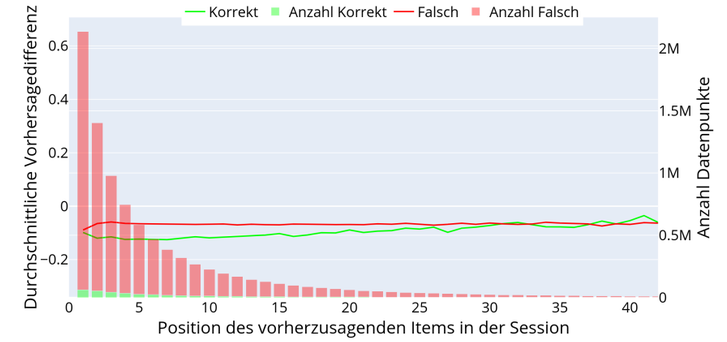

Abbildung 3. Durchschnittliche Vohersagedifferenzen für falsch und korrekt klassifizierte Items pro Position in der jeweiligen Testsession von SR-GNN.

Diskussion und Ausblick

Durch die Verwendung von Vorhersagedifferenzen zwischen dem aktuellen sowie zu vergleichendem Modell konnte festgestellt werden, dass sich das beste Vergleichsmodell (SR-GNN) bei der korrekten Vorhersage am Anfang einer Session sicherer war als das aktuelle Modell. Zudem erzielte dieses beste Modell eine Verbesserung von 69 % (MRR@20) bzw. 44 % (HR@20) zum bestehenden Modell auf dem Testdatensatz mit allen Sessions. Bei den Sessions mit kurzen, langen Sessions sowie bei Betrachtung der Produktkategorien erzielte dieses Modell ebenfalls beste Ergebnisse. Die Ergebnisse basierend auf den Testdatensätzen mit kurzen (langen) Sessions decken sich mit den Erkenntnissen aus [2], wonach längere Sessions bessere Werte für die verwendeten Metriken erzielen als kürzere Sessions.
Anders als bei der Masterarbeit von [1], bei der alle Vorhersagedifferenzen zwischen allen Elementen zwischen zwei Modellen berechnet wurden, konnten für diese Arbeit modell- und ressourcenbedingt die Vorhersagedifferenzen nur für die Schnittmenge an Items von dem jeweiligen Vergleichsmodell und Mahout berechnet werden. Daher wurden diese auf Muster der Vorhersagedifferenzen zwischen korrekten sowie falschen Vorhersagevorgängen untersucht. Für einen Erkenntnisgewinn für das Mahout-Modell könnte es sinnvoll sein, nachvollziehen zu können, warum dieses Modell bei einer gegebenen Eingabe eine Falschklassifikation macht und ein anderes Vergleichsmodell eine korrekte Klassifikation durchführt. Um dies zu realisieren, müssten pro Vorhersagevorgang Vorhersagegewichtungen für alle Items berechnet werden. In dieser Arbeit wurden die Gewichtungen der Items pro Vorhersage in einer Liste mit 20 Items im Arbeitsspeicher gespeichert und nach Beendigung aller Evaluierungen in einer Datei abgelegt. In einer weiterführenden Untersuchung könnten die Gewichtungen aller Items pro Vorhersagevorgang in einer Datei abgespeichert und anschließend auf Muster untersucht werden.
Um zu überprüfen, inwiefern länger zurückliegende Items in einer Session in der Eingabesequenz von SR-GNN einen Einfluss auf Performanz haben, könnte in einer weiterführenden Untersuchung zur Vorhersage des nächsten Items jeweils nur die k-letzten Elemente der Eingabesequenz verwendet werden. Für unterschiedliche k könnte ermittelt werden, ab welcher Länge die Performanz vergleichbar ist mit der Performanz ohne eine Beschränkung der Eingabesequenzlänge. Dies könnte eine Beschleunigung der Vorhersagezeit von SR-GNN mit sich führen, was vorteilhaft für die produktive Umgebung sein könnte, da weniger Ressourcen benötigt werden (Kostenersparnis) und User:innen schneller mit Empfehlungen versorgt werden könnten.

Literatur

[1] Christophe Krech.
Erklärbarkeit maschineller Lernverfahren Feature Engineering mit Black-Box-Modellen.
Master's thesis, Hochschule Darmstadt, April 2019.

[2] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan.
Session-Based Recommendation with Graph Neural Networks.
In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, Honolulu, Hawaii, USA, 2019. AAAI Press.

Referent: Prof. Dr. Markus Döhring

Korreferent: Prof. Dr. Andreas Thümmel