Forschungspraktikum DB-Anwendungsentwicklung - Prognosekombination Potenzialanalyse

Dustin Laudahn
Thomas Olszewski
Amir Rahimi
Nick Scheider
Alexander Volkmann

***Abstract*— Erneuerbare Energien sind in unserem Zeitalter nicht mehr wegzudenken. Um den starken Fluktuationen dieser Art der Energieversorgung entgegenzuwirken, können verschiedenste Prognosemethoden verwendet werden, um den zukünftigen Energieverbrauch in einer Domäne vorherzusagen und dementsprechend handeln zu können. Die Ergebnisse dieser Prognosen können allerdings verschieden stark von den realen Werten abweichen. Um die Differenz zu den realen Werten zu verringern, werden verschiedene Prognosen miteinander kombiniert, um die Stärken der Einzelprognosen zu vereinen. Dafür müssen jedoch geeignete Kombinationsmöglichkeiten gefunden werden. In dieser Potenzialanalyse werden dafür einige Kombinationsmöglichkeiten getestet und gegenübergestellt. Hierbei liefern die Algorithmen des R-Packages „Opera“ und eine Lineare Regression auf Datensätzen bestehend aus Energiewerten die zuverlässigsten Ergebnisse.**

I. Aufgabe und Ziel

Diese Potenzialanalyse geht aus dem ersten Teil des Forschungsprojekt „DB-Anwendungsentwicklung“ hervor. Sie beschäftigt sich mit dem Testen und der Gegenüberstellung von Kombinationsmöglichkeiten für Prognosemodelle, um die Stärken von Einzelprognosen vereinen zu können. Dazu zählen sowohl die Recherche bereits verfügbarer Kombinationsmethoden, als auch die Bewertung der Performance. Anschließend ist es im zweiten Teil das Ziel, diese Methoden in die Prognose Benchmark-Plattform ECAST zu integrieren.

II. Datensätze

Als Grundlage für die Prognosekombination dienen insgesamt vier Datensätze mit Messwerten für den Stromverbrauch und die Stromerzeugung durch sowohl Windkraft als auch Photovoltaik. Diese Datensätze wurden unabhängig voneinander erzeugt und weisen dementsprechend unterschiedliche Beschaffenheiten auf, wie in Tabelle I erkennbar.

III. Durchführung

A. Fehlermaße

Für den Vergleich der Performance von Prognosemodellen und deren Kombination wurde der „Symmetric Mean Absolute Percentage Error“ (SMAPE):

$$SMAPE = \frac{100\%}{N} \sum_{t=1}^{N} \frac{|A_t - F_t|}{(|A_t| - |F_t|)/2} \tag{1}$$

und der „Normalized Root Mean Square Deviation“ (NRMSD) verwendet:

$$NRMSD = 100 \frac{\sqrt{\frac{1}{N} \sum_{t=1}^{N} (A_t - F_t)^2}}{sd(A)} \tag{2}$$

Hierbei steht A_t für den Wert der Vergleichszeitreihe zum Zeitpunkt t und F_t für den Wert der Prognose zum gleichen Zeitpunkt t. Der Wert N ist die Anzahl aller Werte die in Zeitreihe F vorhanden sind. Die Standardabweichung der Vergleichszeitreihe A ist durch $sd(A)$ gegeben.

B. Trainings-/Validierungssplit

Um den Einfluss der Menge an Trainingsdaten für die Kombination zu testen, wurden die verschiedenen Kombinationsmethoden mit einem Trainings-/Validierungssplit von 20%, 40%, 60% und 80% evaluiert. Die Prozentangabe gibt hierbei jeweils die Menge der verwendeten Trainingsdaten an. Dabei wird der Split sowohl für die Kombination, als auch die Einzelprognosen berücksichtigt. Das bedeutet, dass der verwendete Validierungszeitraum für die Berechnung des Fehlers sowohl für die Kombination als auch das Referenzmodell aus den Einzelprognosen gleich ist.

TABLE I
BESCHAFFENHEIT DER DATENSÄTZE

Datensatz	***#Einträge***	**Startdatum**	**Enddatum**	**Granularität**	***#Prognosemodelle***
CH	35.136	01/01/2016	31/12/2016	15min	12
EEM	8.737	01/01/2016	31/03/2016	15min	7
EVA	26.304	01/01/2012	01/01/2015	1h	14
RDS	17.520	01/01/2014	01/01/2016	1h	2

C. Online/Offline Training

Für alle Kombinationsmethoden wird die Möglichkeit des Online und des Offline Trainings getestet. Online bedeutet, dass jeder Tag, für den eine Prognose erstellt wurde, in das Training einbezogen wird, sodass Informationen aus dem Validierungsteil nach und nach in das Modell einfließen. Beim Offline Training werden diese Daten nicht mit in das Training einbezogen, das bedeutet, die Modelle werden nur mittels Informationen der Trainingsdaten erstellt.

D. Kombinationsmethoden

Für die Kombination wird auf eine Reihe verschiedener Methoden gesetzt, um diese gegenüber zustellen. Für das Training und die Validierung werden dafür die vorangehende Splitmethode und die Fehlermaße verwendet und das Training sowohl Offline als auch Online durchgeführt.

1) Support Vector Machine (SVM): Die gegebenen Prognosemodelle werden verwendet, um eine Support Vector Machine zu trainieren, welche dabei aus dem Package „e1071“ [1] stammt.

2) Lineare Regression (LM): Hier werden die Prognosemodelle verwendet, um eine Lineare Regression durchzuführen, die aus dem Package „Caret“ [2] stammt.

3) Mean Best_n (MBn): Die Einzelprognosen werden aufsteigend nach Fehler sortiert, dabei wahlweise nach SMAPE oder NRMSD. Anschließend wird der Durchschnitt pro Eintrag über die *n* besten Methoden errechnet.

4) Opera: Bei dem „Opera“ Package handelt es sich um ein R-Package, welches auf den Anwendungszweck der Prognosekombination abzielt. Dabei werden die zur Verfügung gestellten Algorithmen EWA, FS, Ridge, MLpol und OGD verwendet. [3], [4]

IV. AUSWERTUNG DER DATEN

a) MBn: Das Verfahren des Durchschnitts der *n* besten Methoden konnte nur in seltenen Fällen die beste Einzelprognose schlagen. Stattdessen war der Fehler

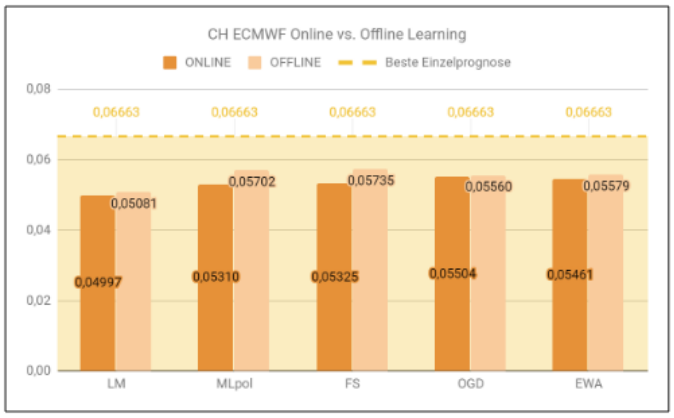

Fig. 1. Vergleich SMAPE zwischen Online und Offline Training über 80% Trainingsdaten auf dem CH ECMWF Datensatz

beim Großteil der Ergebnisse bis zu doppelt so hoch wie der des Referenzmodells. Für einen Faktor von $n = 2$ werden die besten Ergebnisse erzielt.

b) SVM: Die Support Vector Machine liefert sehr fluktuierende Ergebnisse und erweist sich in einigen Fällen als der größte Ausreißer. Weiterhin ist die Berechnungszeit mit mehreren Stunden sehr groß, weshalb eine weitere Verfolgung des Ansatzes verworfen wurde.

c) LM: Die Lineare Regression weist zusammen mit den Funktionen des Opera Packages die besten Ergebnisse auf. Sie schneidet konstant besser als das Mean Best_n Verfahren ab und weist in den meisten Fällen einen niedrigeren Fehler auf, als das Opera Package, wie in Grafik 1 zu erkennen ist.

d) Opera: Die Funktionen des Opera Packages schaffen es in den meisten Fällen annähernd die Performance der besten Einzelprognose zu erreichen oder diese sogar zu überbieten, wie auch in den Tabellen III und II erkennbar ist.

e) Training-/Validierungssplit: Für den Training-/Validierungssplit lässt sich keine direkte Aussage formulieren, da dieser von der Beschaffenheit des Datensatzes abhängt. Wie in Tabelle I ersichtlich ist, schwankt diese bezüglich der Menge an Dateneinträ-

TABLE II

AUSWERTUNG DER KOMBINATIONEN MIT NRMSD FEHLER NACH 60% TRAININGSDATEN UND 40% VALIDIERUNGSDATEN

	Referenz	MLpol	FS	OGD	EWA	LM	MBn, n=2	SVM	Ridge
CH_ECMWF	23	21,9	21,9	22,5	21,5	**19,8**	38,5	20,1	21,9
EVA	65	63,4	67,8	64,1	63,4	59,3	**55,4**	102,8	63,4
EEM	39,9	45,7	52,8	45,8	45,5	**45,4**	60,8	**45,4**	45,6
RDS	20,7	19,3	**19,1**	19,2	19,3	**19,1**	26	19,2	19,4

TABLE III

AUSWERTUNG DER KOMBINATIONEN MIT SMAPE FEHLER NACH 60% TRAININGSDATEN UND 40% VALIDIERUNGSDATEN

	Referenz	MLpol	FS	OGD	EWA	LM	MBn, n=2	SVM
CH_ECMWF	6,66%	**5,22%**	5,49%	5,35%	5,58%	5,42%	6,33%	- *
EVA	87,32%	60,74%	**59,92%**	60,12%	60,83%	124%	119,85%	133,89%
EEM	33%	37,02%	42,81%	38,09%	37,02%	**36,7%**	52,03%	40,11%
RDS	155,35%	161,48%	161,46%	160,76%	161,47%	**159,4%**	162,63%	159,47%

*Berechnungszeit des SVM-Werts zu groß

gen. Abgesehen von dem EEM-Datensatz, welcher mit Abstand die wenigstens Datenpunkte aufweist, führte ein Training über 20% der Daten und somit mit 80% der Daten zur Validierung zu den besten Ergebnissen, während für andere Datensätze mit 60% Trainings- und damit 40% Validierungsdaten am besten performt wurde.

f) Online vs. Offline Training: Beim Vergleich des Online gegenüber des Offline Trainings, zeigt sich, dass das Online Training bessere Ergebnisse liefert. Grafik 1 zeigt, dass für den CH Datensatz über 80% Trainingsdaten dank Online Training abermals eine Verbesserung erreicht werden konnte gegenüber dem Offline Training.

V. ZUSAMMENFASSUNG UND AUSBLICK

Während der Analyse von Prognosekombinationsmethoden stellten wir fest, dass es keinen statischen Ablauf gibt, um garantiert optimale oder gar positive Ergebnisse zu erzielen. Die Ergebnisse hängen stark mit der Art und der Größe des Datensatzes zusammen. So stellen wir besonders beim kleinsten Datensatz, EEM, je nach Wahl des Trainingszeitraums große Schwankungen bei den Ergebnissen der Kombination fest. Besonders wichtig ist es für Datensätze mit Saisonalität mindestens einen vollen Zyklus der Saison zum Training zu verwenden um sinnvolle Ergebnisse in der Vorhersage zu erhalten. Um die Ergebnisse der Kombinationen vergleichen zu können sollte darauf geachtet werden, dass die Kombinationen auf Datensätzen gleichen Typs erstellt werden. Außerdem sollten neben dem SMAPE und dem NRMSD weitere Fehlermaße untersucht werden um einen besseren Überblick über den Einfluss dieser zu bekommen. Abschließend lässt sich sagen, dass die hier untersuchten Methoden des Opera Packages und der Linearen Regression in der Lage sind bereits vorhandene Prognosemodelle zu kombinieren, um verbesserte Ergebnisse zu erhalten. Eventuell sollten noch weitere Machine Learning Modelle untersucht werden, die noch bessere Kombination bestimmen könnten. So wäre ein weiterer Ansatz der in Betracht gezogen werden könnte, das Lernen einer optimalen Gewichtung anhand von Ground-Truth Gewichten mit Hilfe eines Supervised-Learning Algorithmus.

REFERENCES

[1] e1071 R-Package with LM, https://cran.r-project.org/web/packages/e1071/e1071.pdf

[2] Caret R-Package, https://cran.r-project.org/web/packages/caret/caret.pdf

[3] Opera R-Package, https://cran.r-project.org/web/packages/opera/opera.pdf

[4] Goude, Yannig & Gaillard, Pierre. (2016). OPERA, a R package for online aggregation of experts. https://www.researchgate.net/publication/304678996_OPERA_a_R_package_for_online_aggregation_of_experts