StatP_Ensembles_Exposé

Juan Sebastian Aristizabal Ortiz und Tchoumkeu Ngatat Emmanuel

11/17/2021

Using ensembles to improve forecast performance

Problembeschreibung

Allgemein dienen die Methoden der deskriptiven Statistik dem besseren Verständnis und Veranschaulichung von Ergebnissen.

Öfters ist es unerheblich, ob es sich um die Ergebnisse einer individuellen Modellsimulation oder die weitere Auswertung mehrerer Simulationen zusammen handelt. > Simulation?

Modellensembles beschreiben eine Bandbreite möglicher Zukünfte, indem sie die Reaktion auf eine Bandbreite verschiedener Szenarien simulieren, helfen uns deren Unsicherheiten zu verstehen und übertreffen so einzelne Modelle und Prognostiker in Bezug auf die Vorhersageleistung. > ? Simulation

Empirisch hat es sich jedoch als überraschend schwierig erwiesen, einfache Mittel- oder Median-Ensembles zu verbessern. > Nicht verbessern sondern auszuwählen.

Daher ist von relevantem Interesse zu verstehen welche Form des Ensembles, je nach Datentypen genutzt werden soll, um Prognosen in einem angewandten Vorhersageumfeld signifikant zu verbessern. > Was meinst du mit angewandte Vorhersageumfeld?

Datenbeschreibung

Zur Analyse stehen Daten aus einem Prognosen-challenge, das zu Beginn der COVID-19 Zeiten in der UK durchgeführt wurde. > Ein Datensatzt bezieht sich auf die UK. Den Hauptdatensatzt beinhaltet mehreren Ländern.

Alle Vorhersagen werden in einem quantilbasierten Format erfasst. Das heißt, dass die Prognostiker eine Vorhersageverteilung in Form von 23 Quantilen (11 Vorhersageintervalle plus der Medianvorhersage) bereitstellen, die angeben, wie wahrscheinlich sie glauben, dass der wahre beobachtete Wert in einen bestimmten Bereich fällt.

Wir möchten untersuchen inwiefern sich die Aggregation menschlicher Prognosen von der Aggregation modellbasierter Prognosen unterscheidet und vergleichen deshalb Resultaten der Modellensembles angewendet auf den Prognosen-challenge mit Prognosen, die von verschiedenen Forschungseinrichtungen an den European Forecast Hub übermittelt wurden. > Wir haben einmal ein Datensatzt, die aus menschlichen Vorhersagen besteht und einer, der aus Modelvorhersagen besteht.

Mithilfe von den wahren beobachteten Werte (auch verfügbar beim European Forecast Hub), bilden wir dann score-Funktionen, welche auf einer flexiblen Weise wiedergeben sollen, wie geeignet die jeweiligen Methoden tatsächlich sind.

Zielbeschreibung

Diese Arbeit soll sich mit verschiedenen Ensemblemethoden befassen und daraus wichtige Schlussfolgerungen hervorbringen bezüglich der Qualität, der Präzision und der geeigneten Form des Ensembles, das empirisch zu wählen ist.

Wir untersuchen wie sehr einzelne Modelle zu einem Ensemble beitragen. Wir identifizieren Situationen, in denen das Hinzufügen eines Modells von Vorteil ist oder nicht und treffen schließlich Annahmen über die Stabilät der Prognoseleistung verschiedener Ensemble-Typen.

Methodenbeschreibung

Kern unseren Betrachtungen werden folgende Ensemble-Typen sein:

- Einfaches Mean-Ensemble (untrainiert)
- Median Mean-Ensemble (untrainiert)
- Weighted mean ensemble (trainiert)
- Weighted median ensemble (trainiert)

Hierbei steht 'trainiert' im Klammer für Modelle, die Daten aus der Vergangenheit zur Hilfe nehmen und 'untrainiert' für Ensemble-Typen, die ohne Hilfdatensätze durchlaufen. > nicht ganz. Lass das besprechen.

Implementierung

Wir spalten die verschiedenen Herangehensweisen in vier teilen:

- Zuerst bewerten wir die Leistung in Abhängigkeit von der Anzahl der Mitgliedermodelle (n). Iterativ werden n Modelle zusammen zu einem Ensemble aggregiert. Zur Auswertung werden die Ergebnisse aller möglichen Kombinationen dieser n Modelle gemittelt und mit der durchschnittlichen Leistung vergliechen.
- Anschließend analysieren wir die Leistung eines Ensembles in Abhängigkeit von der Ähnlichkeit seiner Mitgliedsmodelle. Für jedes Ensemble berechnen wir eine Ähnlichkeitsbewertung mithilfe der (paarweise) Cramér-Distanz (Cook): und überprüfen, ob Ensembles mit einem höheren Durchschnittsabstand besser abschneiden.
- Wir betrachten dann die Auswirkungen der Hinzunahme eines Modells (auch wenn dieses verzerrt ist) zu einem Ensemble und untersuchen diese Auswirkungen für verschiedene Ensemble-Typen.
- Schließlich evaluieren wir wie stabil (also robust) die verschiedenen Ensemble-Typen sind. Das heißt, dass wir durch Minierung der absoluten Differenzen überprüfen, ob Beobachtungen mit grösseren Abweichungen durch Quadrierung ein relativ niedrigeres Gewicht erhalten (in dem Fall wäre der Ensemble weniger anfällig für Ausreißer)

Test-Framework

Die Beurteilung von Ergbnissen soll größtenteils mithilfe von selbst-geschriebene score-Funktionen erfolgen.

Prognosen in einer quantilbasierten Prognose können mit dem gewichteten Intervall-Score (WIS, 'weighted interval score') bewertet werden: je niedriger desto besser. Der WIS wird in drei Komponenten zerlegt: WIS = Streuung + Übervorhersage + Untervorhersage. Der WIS steht außerdem in engem Zusammenhang mit dem absoluten Fehler. Über- und Unterprognosen sind daher auch als eine Form des absoluten Prognosefehlers zu verstehen.

Prognosenkalibrierung

Im Durchschnitt sollten alle 50% oder 90% Vorhersageintervalle idealerweise 50% oder 90% der wahren Beobachtungen abdecken. Durch den Vergleich der nominalen und empirischen Abdeckung können wir dann feststellen, ob ein Modell über- oder untersicher ist. Mittels verschiedene räumliche Statistik-methoden PIT-Histogramme oder Geometrische Darstellungen) können zusätzlich benutzt werden, soimulierung Prognosen in R .