supper&supper

Ertragsvorhersage von Hybriden

Projektziel

Projektziel war es eine Prognose zu erstellen, wie sich hybrides Getreide an neuen Standorten unter variierenden Umweltbedingungen verhält. Basierend auf den bereitgestellten Daten wurde ein Modell kreiert, um die Getreideerträge vorherzusagen und anschließend auf neue Kombinationsmöglichkeiten neuer Hybridgetreide und Hybridorte angewandt.

Bereitgestellte Daten

Die folgenden Datensätze wurden bereitgestellt: Ein Datensatz beinhaltete Informationen über eine Vielzahl von Hybriden, ihren Ertrag und Standort über den Zeitraum 2001 - 2016, sowie eine Vergleichsspezies. Ein anderer Datensatz umfasste die Hybride und Standorte für das Jahr 2017, für welche die Erträge vorherzusagen waren. Ein weiterer Datensatz beinhaltete genetische Marker aller Hybride und Bodenparameter, in Form von monatlichen Werten für den gesamten Zeitraum und alle Standorte. Insgesamt bildeten diese Daten eine Zeitspanne von 15 Jahren für mehr als 2.000 Hybrid-typen und Standorte ab.

Herausforderungen

Die bereitgestellten Daten mussten analysiert und bereinigt werden, um sicherzustellen, dass die Data Mining-Algorithmen richtig funktionieren. Das Zusammenfügen der Daten brachte eine Matrix hervor, die mehr als 3 Milliarden Datenpunkte enthielt. Wir mussten daher einen Weg finden, diese Datenpunkte zu reduzieren, um unsere Algorithmen anwenden zu können ohne große CPU-Rechenleistung zu benötigen. Schließlich mussten Annahmen für ein Jahr getroffen werden, für das noch keine Wetterparameter verfügbar waren. Dies machte es notwendig, ein zweites VorhersageModell für das Wetter zu erstellen.

Angewandte Methoden

Datenbereinigung

Für jeden zur Verfügung gestellten Datensatz wurde eine gründliche Analyse der Instanzen durchgeführt. Ausreißer innerhalb verschiedener Standorte wurden durch überlappende geographische Daten, Klimainformationen und eine Vielzahl von Events identifiziert. Die Ausreißer wurden dann entfernt.

Big Data Problem

Viele Instanzen hatten eine große Schnittmenge bezüglich des genetischen Materials. Deshalb wurde die Dimensionalität des Datensatzes signifikant reduziert, indem eine Multi-Faktor-Analyse ohne großen Informationsverlust verwendet wurde.

Wettervorhersage

Ein Autoregressive Integrated Moving Average Modell (ARIMA) wurde den Residuen des Fourier-Regressions-Modells angepasst. Ein Wave-Type Kovarianz-Modell wurde für die zeitliche Dimension, ein exponentielles KovarianzModell wurde für die räumliche Dimension angewandt. Für die Vorhersage der Raum-Zeit-Zufallsfelder wurde ein räumlich-zeitliches Kriging angewandt. Das Wetter wurde mit einer Genauigkeit von 95 % vorhergesagt.

Ertragsvorhersage

Der reduzierte Datensatz des genetischen Materials, die Wetter-, Boden- und Ertragsdaten wurden in einen Datensatz zusammengefasst und als Trainingssatz für das Modell (3-fache Cross-Validation) verwendet. Einige Algorithmen wurden verglichen: Künstliche Neuronale Netzwerke, Support-Vektor-Regression und Decision Trees. Die beste Vorhersagegüte wurde mit einem Random Forest Modell erreicht. Die Hybrid-Leistung wurde so mit einer Genauigkeit von 75 % vorausgesagt.

Projektergebnis

Das Modell wurde erfolgreich für die neuen Hybride an 20.000 neuen Standorten angewandt. Damit konnten wir die leistungsfähigsten Spezies für das Jahr 2017 identifizieren.