FoPra: Maschinelles Lernen zur Klassifizierung von Sequenzdaten

Kien Nguyen
Nick Lowack

June 2021

1 Kurzbeschreibung

Im Rahmen des Fortgeschrittenen Praktikums wurde ein Programm entwickelt, welches mit Maschine Learning klassifiziert ob ein Scaffold seinen Ursprung im Zellkern hat oder nicht.

2 Entwicklungsprozess

Uns wurde ein Datensatz zur Verfügung gestellt, welcher die sequenzierten Scaffolds mit einigen Kenndaten wie z.B. tRNA Wert enthält. Als nächstes mussten wir uns entscheiden in welcher Sprache und Entwicklungsumgebung wir uns dem Problem nähern wollten.

Nach Absprache mit unserem Betreuer, Roman Martin entschieden wir uns für Python, da wir als Team bereits Erfahrung mit der Maschine Learning Bibliothek scikit-learn gesammelt haben.

Ein Problem, welches uns durch den gesamten Entwicklungsprozess begleiten sollte, war die Tatsache das wir mit unbalancierte Daten arbeite mussten. Diese standen in einem Verhältnis von 99,63 % mit Ergebnis 0 und 0,37 % mit Ergebnis 1

3 Zeitplan

Unser Zeitplan gliedert sich entlang des Meilensteinplans für unser Projekt.
4.5.2021 - 25.5.2021:
Die ersten 3 Wochen sollten dafür genutzt werden unser generelles Wissen über Maschine Learning und die benötigten Python Bibliotheken zu erweitern. Auch haben wir eine allgemeine Datenanalyse unseres Datensatzes vorgenommen um etwaigen Problemen vorzubeugen.
26.5.2021 - 15.6.2021:
Dieser Zeitraum sollte dafür genutzt werden den biologischen Hintergrund der Aufgabenstellung besser zu verstehen. Außerdem war das Ziel bis zum nächsten Meilenstein einen erstes Modell zu Klassifizierung der Rohdaten zu erstellen und zu evaluieren.
16.6.2021 - 22.6.2021:
bis zu diesem Meilenstein sollte die Ausarbeitung verschiedener Modelle vorgenommen werden. Auch sollten diese Gegeneinander evaluiert und bewertet werden. Im Falle von enttäuschenden Ergebnissen sollten neue Daten generiert werden.
23.6.2021 - 29.6.2021:
In diesem Zeitabschnitt sollte die Vorbereitung auf die Endabnahme vorgenommen werden. Die Dokumentation und Präsentation sollten vorbereitet werden, als auch eine zusammenfassende Datenpräsentation aus allen Modellen erstellt werden.

4 Definition des Zielsystems

Das System ist ein Python Programm welches einen Datensatz von sequenzierten Scaffolds bekommt. Das System bereinigt die Daten und teilt sie in Trainings- und Testdatensatz ein. Mit verschiedenen Maschine Learning Klassifikationsverfahren werden die Daten nach ihrem Ursprung aus dem Zellkern eingeordnet. Schließlich werden die Klassifikationsverfahren nach ihrem F1-Score bewertet.

5 funktionale Anforderungen

Da das System für die Forschung entwickelt wurde ist die einzige Funktionale Anforderung die Einordnung der Sequenzdaten in die Kategorie 0 oder 1.

6 Nicht funktionale Anforderungen

Die Bedienbarkeit des System ist für Menschen mit Programmiererfahrung sehr einfach, da nur ein neuer Datensatz eingespeist werden muss und die Pipeline ggf. verändert werden kann. Außerdem ist der Code ausreichend Dokumentiert und die Variablen haben passende Namen. Dies spiegelt sich auch in dem Aussehen und der Handhabung des Programms wieder.

Zur Zuverlässigkeit ist zu sagen, dass das System mit jeder neuen Ausführungen tendenziell andere Ergebnisse erzielt aber eine gewisse Ordnung stets vorhanden ist.

Leistung und Effizienz sind in dem Sinne gut, das keine ineffizienten Algorithmen oder überflüssiger Code vorhanden sind. Die Maschine Learning Klassifikationsverfahren können jedoch je nach Verfahren länger brauchen.

7 Entwurf

Es lässt sich feststellen, dass das System einer linearen Struktur folgt in der ein Schritt nach dem nächsten ausgeführt wird.

Der erste Schritt war die Einordnung des Projekts bzw. eine möglichst genaue Definition der Anforderungen.

Als nächstes konzentrieren wir uns auf die gegebenen Daten und lassen uns die Charakteristiken dieser Visualisieren. Hierbei lassen wir uns die z.B. die Ausreißer anzeigen oder untersuchen Korrelationen zwischen Attributen.

Nun werden die Daten prepariert. So werden beispielsweise die Ausreißer entfernt, solange sie den Wert [Organelle = 0] besitzen und die übrigen Werte werden standardisiert

der nächste Schritt den wir vorgenommen haben war die Erstellung einer Shortlist von vielversprechenden Kategorisierungsverfahren. wir nutzen ein Kreuzvalidierungsverfahren um die Daten aufzuteilen und messen die Perfomance der einzelnen Modelle.

Nachfolgend filtern wir die Shortlist nach den besten Modellen und feinjustieren die Hyperparameter. Dann nutzen wir für alle Top Modelle einmal grid search und einmal random search. Der endgültige F1-score der Modelle wir ausgegeben.

8 Technologien

Die benutzte Programmiersprache ist Python. Für die Maschine Learnig Prozesse nutzen wir hauptsächlich die scikit learn Bibliothek. Die Datenrepräsentation wird mit Panda erledigt und mathematische Prozesse mit der numpy Bibliothek.

9 Bewertung

Ein endgültige Bewertung des Systems gestaltet sich schwierig. Dadurch das die Daten aus echten Forschungsergebnissen stammen gestaltet, ergeben sich sehr unbalancierte Daten. Außerdem haben diese Daten in den meisten Fällen sehr viele null Einträge in den Feldern.

Es zeigt sich das selbst unsere besten Modelle mit feinjustierung der Hyperparameter selten einen F1-score von über 0.5 erreichen. Grundsätzlich sollten wir, im Rahmen unserer Möglichkeiten alles getan haben um ein gutes Modell zu generieren aber vielleicht reicht die Datenlage einfach nicht aus.