Titel: Visualisierung eines Suchraums für Symbolische Regression.

Einleitung: Symbolische Regression ist ein spezieller Ansatz der Regressionsanalyse und damit ein Ansatz des überwachten Lernens. Gegeben ist eine Menge an Beobachtungen von mehreren numerischen Variablen. Gesucht ist eine Formel um die Werte einer ausgewählten Variable aus den bekannten Werten der anderen Variablen zu prognostizieren. Weitere Parameter sind die Menge der mathematischen Operatoren und Funktionen die innerhalb der Formel verwendet werden dürfen sowie eine maximale Länge für die Formel.

Der Fokus dieser Arbeit ist die genauere Analyse des Suchraums für Symbolische Regression. Dazu generieren wir mit einem sehr einfachen Backtracking-Ansatz alle möglichen Formeln die sich aus einer kontext-freien Grammatik (siehe Abb. 1) für eine formale Sprache ableiten lassen und eine vorgegebene maximale Länge nicht überschreiten. Diese Menge aller Formeln repräsentiert alle möglichen Lösungen und kann einmalig generiert werden da sie unabhängig von der gegebenen Datenmenge ist. Erst in Kombination mit Daten kann der Prognosefehler aller Formeln berechnet werden und die optimale Formel ausgewählt werden. Die Hypothese dieser Arbeit ist, dass die Formeln einmalig anhand Ihrer Ähnlichkeit kartographiert werden können und dass es dadurch möglich ist für eine gegebene Datenmenge effizienter die optimale Formel zu finden.

```
G(Expr):
Expr       -> "const" "*" Term "+" Expr | "const" "*" Term "+" "const"
Term       -> RecurringFactors "*" Term | RecurringFactors | OneTimeFactors

RecurringFactors -> VarFactor | LogFactor | ExpFactor | SinFactor
VarFactor  -> <variable>
LogFactor  -> "log" "(" SimpleExpr ")"
ExpFactor  -> "exp" "(" "const" "*" SimpleTerm ")"
SinFactor  -> "sin" "(" SimpleExpr ")"

OneTimeFactors -> InvFactor "*" SqrtFactor "*" CbrtFactor |
                  InvFactor "*" SqrtFactor               |
                  InvFactor "*"            CbrtFactor    |
                          SqrtFactor "*" CbrtFactor     |
                  InvFactor  |
                  SqrtFactor |
                  CbrtFactor
InvFactor  -> "1/" "(" InvExpr ")"
SqrtFactor -> "sqrt" "(" SimpleExpr ")"
CbrtFactor -> "cbrt" "(" SimpleExpr ")"

SimpleExpr -> "const" "*" SimpleTerm "+" SimpleExpr  | "const" "*" SimpleTerm "+" "const"
SimpleTerm -> VarFactor "*" SimpleTerm | VarFactor

InvExpr -> "const" "*" InvTerm "+" InvExpr | "const" "*" InvTerm "+" "const"
InvTerm -> RecurringFactors "*" InvTerm  |
           RecurringFactors "*" SqrtFactor "*" CbrtFactor |
           RecurringFactors "*" SqrtFactor               |
           RecurringFactors "*"            CbrtFactor    |
                           SqrtFactor "*" CbrtFactor     |
           RecurringFactors |
           SqrtFactor       |
           CbrtFactor
```

Abbildung 1: Kontext-freie Grammatik für Symbolische Regression

Methoden: Wir generieren alle Formeln mit zwei unterschiedlichen Variable (x,y) und beschränken die Länge der Formeln auf fünf Variablen. Im nächsten Schritt werden drei

Datensets mit 1000 x,y-Paaren systematisch erzeugt. D1= $\{ (x,y) \mid x,y \in [-3 \dots 3] \}, D2 = \{(x,y)|x,y \in [0 \dots 3]\}, D3 = \{(x,y) |x,y \in [-3 \dots 0]\}$. Für jede Formel wird das Funktionsergebnis für diese drei Datenmengen berechnet. Formeln mit ungültigen Werten (undefiniert oder unendlich) sind unerwünscht und werden ausgefiltert. Die übrigen Funktionen werden mithilfe UMAP [1] vom ursprünglichen 1000-dimensionalen Raum auf einen 2-dimensionalen Raum zugeordnet. Bei dieser Dimensionsreduktion bleiben Nachbarschafts-Beziehungen erhalten; ähnliche Funktion im ursprünglichen Raum werden im niedrig-dimensionalen Raum näher zueinander dargestellt.

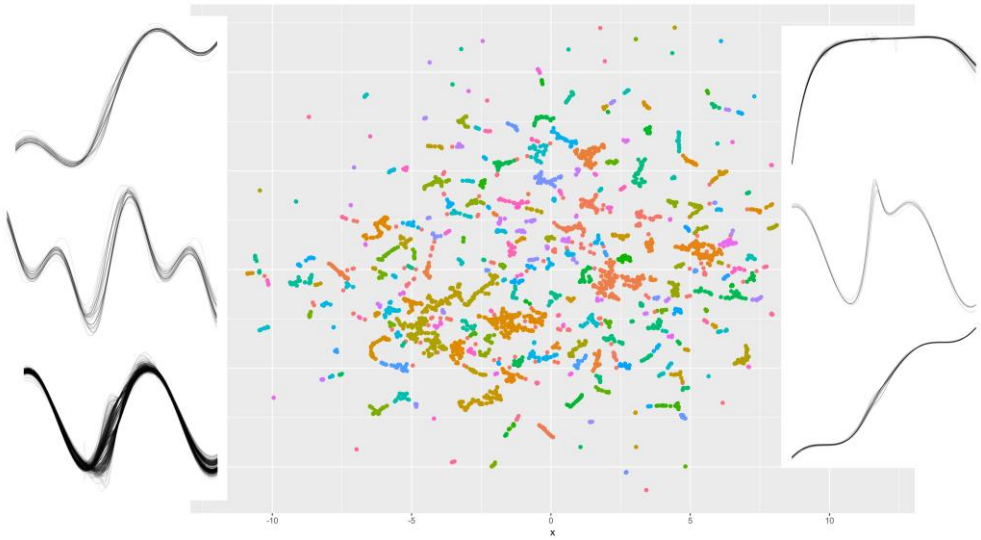

Abbildung 2: Visualisierung des Lösungsraums. Jeder Punkt repräsentiert eine Funktion. Benachbarte Funktionen sind ähnliche Funktionen. Die Funktionen wurden nach Ähnlichkeit gruppiert (Farbe=Gruppe). An Rand sind sechs ausgewählte Cluster dargestellt.

Ergebnisse: Es konnten 7.725.938 unterschiedliche Funktionen aus der Grammatik erzeugt werden. Dabei wurden bereits äquivalente Darstellungen erkannt und gefiltert. Nach der Evaluierung verbleiben für D1 84.381 gültige Funktionen, für D2 599.015 für D3 93.660. Insgesamt verbleiben777.057 Funktionen. Durch die Grammatik und die Erkennung äquivalenter Darstellungen konnte die Menge der möglichen Lösungen im Vergleich zu herkömmlichen Verfahren stark reduziert werden. Durch die Filterung der unerwünschten nicht-stetigen Funktionen konnte der Suchraum weiter um den Faktor 10 reduziert werden. Das Ergebnis ist der Dimensionsreduktion ist in Abb. 2 dargestellt. Es sind klar definierte Cluster ähnlicher Funktionen erkennbar. Diese Cluster-Struktur und die Nachbarschaftsbeziehungen könnten in einem Suchverfahren genutzt werden um für ein gegebenes Datenset innerhalb weniger Schritte zu einer nahe-optimalen Lösung für symbolische Regression zu finden.

Quellen:

[1] McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018