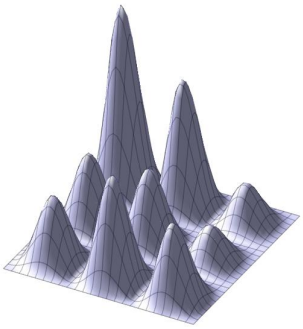
Grundlagen und Einsatz Genetischer Algorithmen

Eckhard Reh

Prinzip

Zur Optimierung chemischer Prozesse oder Messungen wird oft die Simplex- oder Response-Surface-Modelling-Strategie eingesetzt, die aber nicht erfolgreich sind z.B. bei Nebenoptima.



Bei komplexen, hochdimensionalen Zusammenhängen wie in Abbildung 1 skizziert, bieten genetische Algorithmen zur Lösung von Optimierungsproblemen und zur Extremwertsuche eine gute Option bei der Auffindung des globalen Optimums. Genetische Algorithmen, GA, sind heuristische Suchalgorithmen für eine numerische Näherung, die in vielerlei Bereichen zur Anwendung kommen, z.B. Auffindung kleinster Wegstrecken in der Logistik, Abarbeitung einer Probewarteschlange im Labor oder Maximierung eines Response-Signals bei mehreren Variablen in der Chemie.

Abb. 1: Veranschaulichung komplexer Prozesse mit diversen Maxima

Die GA basiert analog zur biologischen Evolutionsstrategie auf [1, 2]
Population, Individuum, Chromosom, Gen und Fitness bzw.
Wachstums-Zyklen mit Kreuzung, Mutation und Reproduktion.
Die relevante Eigenschaft eines Individuums wird als Fitness (Zielgröße) bezeichnet, diese wird bestimmt durch sein Chromosom (Faktor) mit dazu gehörigen Genen.
Mehrere Individuen bilden die Population.
Ein neues Individuum entsteht durch Kreuzung und evtl. Mutation, eine neue Population resultiert abschließend aus der Reproduktion der aktuellen Individuen (Wachstums-Zyklus).
Diese Abfolge wird vielfach wiederholt. Anders als bei der biologischen Evolution gibt es bei der numerischen Anwendung kein Wachstum der Population, hier werden bei konstanten Individuenzahlen deren Eigenschaften (Fitness) von Zyklus zu Zyklus verbessert.

Als einfachste Anwendung z.B. einer Optimum-Suche sei ein eindimensionales Beispiel mit einem Maximum aufgeführt.

Gesucht wird das maximale Signal, Zielgröße y [V] (Fitness) aus Untersuchungen des Prozesses bei der Optimierung. Die Fitness hängt ab von den aktuellen Variablen, hier von Faktor x_1 (Chromosom), z.B. im Masse-Bereich von 0 bis 50 [g].
In der GA-Implementierung wird der Faktor nicht als Realgröße eingesetzt sondern als Binärzahl kodiert, d.h. für Faktor 20 g steht 010100.
(Kodierung auf x_1-Intervall [0, 50] mittels 6-Bit; für Fitness gelte $y = 75 - 0{,}12\ (x_1 - 25)^2$, d.h. Parabel mit Maximum bei $x_1 = 25$)

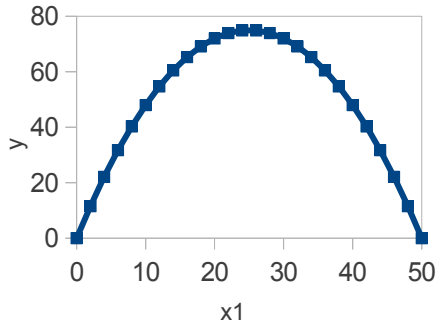


Abb. 2: Einfaches GA-Optimierungsbeispiel

Die Zuordnung zu den biologischen Termini ist in Abbildung 3 dargestellt:

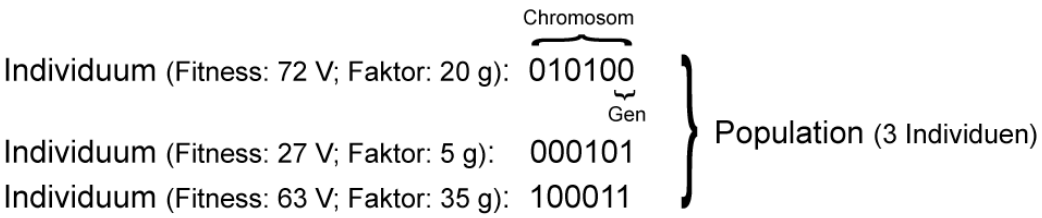


Abb. 3: Verdeutlichung Gen, Chromosom bzw. Individuum, Population bei GA

Die Binärzahl-Kodierung ist entsprechend der Aufgabenstellung anzupassen.
Liegen mehrdimensionale Prozesse z.B. mit 3 Faktoren (Variablen x_1, x_2, x_3) vor, werden 3 Chromosome pro Individuum definiert, die unabhängig voneinander einem GA-Zyklus unterzogen werden.
Mithilfe arithmetischer Transformationen (z.B. Multiplikations-Faktor 1000) können mit der Binärkodierung auch reelle Zahlen behandelt werden.

GA-Zyklus
Bei der Initialisierung werden jedem Individuum zufallsgeneriert Startgene zugewiesen. Es folgt die wiederholte Ausführung eines Zyklus bestehend aus Kreuzung, Mutation, Reproduktion und Identifizierung der beiden Individuen mit den bislang besten Fitness-Werten. Die Zyklen werden ausgeführt bis ein Abbruch-Kriterium erfüllt ist.

Abbildung 4 verdeutlicht den Prozess.

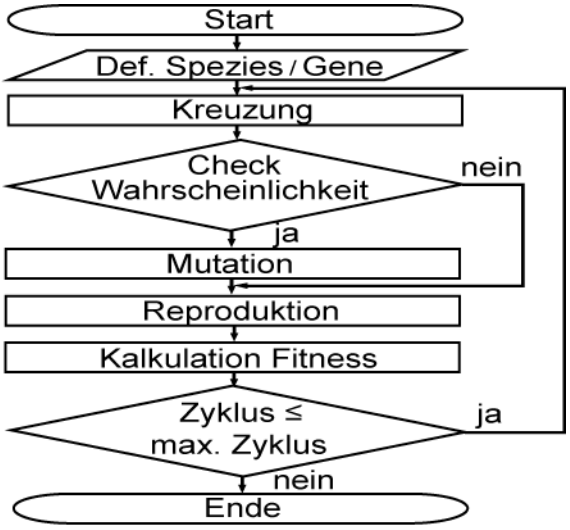


Abb. 4: Prozess Genetischer Algorithmus

Kreuzung
Die Kreuzung (Rekombination) erfolgt am besten bzw. zweitbesten Individuum der letzten Population (Vater, Mutter). Der Kreuzungspunkt des Chromosoms wird zufallsgeneriert festgelegt, es entstehen 2 neue Individuen (a, b). Die Kreuzung wird gleichzeitig und unabhängig an jedem Chromosom durchgeführt. In das erste neue Individuum werden bis zum Kreuzungspunkt die Gene des Vater-Individuums übertragen danach die Gene des Mutter-Individuums, für das zweite Individuum umgekehrt. Z.B.:

	↓ Kreuzungspunkt		
Vater-Individuum:	010100		Individuum a: 010011
Mutter-Individuum:	100011	⇒	Individuum b: 100100

(Alternativ kann die Kreuzung auch zwischen 2 zufällig gewählten Individuen erfolgen.)

Mutation
Bei der Mutation wird an einem zufallsgeneriert festgelegten Mutationspunkt im Chromosom der Gen-Eintrag umgekehrt, dies erfolgt beim bestem bzw. zweitbesten Individuum der aktuellen Population. Sie geschieht gleichzeitig, unabhängig an jedem Chromosom. Die Mutation erfolgt nicht in jedem Zyklus sondern abhängig von einer vorgegebenen Wahrscheinlichkeit.

	↓ Mutationspunkt		
bestes Individuum:	010100	⇒	010000
zweitbestes Individuum:	100011	⇒	100111

(Alternativ kann die Mutation auch an 2 zufällig gewählten Individuen erfolgen.)

Reproduktion
Bei der biologischen Evolution überleben eine Reproduktion (Selektion) bevorzugt Individuen mit der höheren Fitness, analog auch bei der numerischen Implementierung. Nach Kreuzung und Mutation wird das Chromosom des Individuum mit dem schlechtesten Fitness-Wert ersetzt durch das Chromosom des Individuums mit zweitbestem Fitness-Wert der aktuellen Population.

(Alternative ist die Tournament-Reproduktion. Es werden zufällig 2 Individuen gewählt, das mit der besseren Fitness wird in die neue Population übernommen. Dies wird entsprechend der Individuenzahl wiederholt.)

Charakteristika
Folgende Parameter sind vom Anwender vorzugeben:
Populationsgröße, Chromsomzahl, Genzahl, Mutations-Wahrscheinlichkeit, Stop-Kriterium.

Populationsgröße
Sinnvollerweise sollte bei mehreren Nebenoptima der Prozess von diversen Positionen gestartet werden, um nicht auf dem gleichen Nebenoptimum zu enden. Bei dem GA-Algorithmus starten bei 3 Individuen in der Population diese gleichzeitig von 3 zufällig festgelegten Positionen.
Die Zahl der Individuen kann beliebig erhöht werden mit den Konsequenzen, dass das globale Optimum mit größerer Sicherheit ermittelt wird, gleichzeitig der Aufwand für die Untersuchungen und die Rechenzeit linear zunimmt.

Chromsom-Zahl
Für jeden Faktor (x_1, x_2, ... x_p) des Optimierungsproblems wird ein eigenes Chromosom als Binärzahl generiert. Die Kreuzung erfolgt pro Zyklus für jedes Chromosom bei identischem Kreuzungspunkt, die Mutation nur an einem zufällig gewählten Chromosom.

Gen-Zahl
Die Zahl der Gene wird durch die Aufgabenstellung vorgegeben. Soll z.B. in obigem Fall der Faktor x_1 mit 1 Nachkommastelle verwendet werden, kann dieser mit 10 multipliziert und dann binär kodiert werden. Bei einem Wertebereich von 0,0 bis 50,0 [g] würden Chromosome mit 9 Genen zum Einsatz kommen.

Mutations-Wahrscheinlichkeit
Je größer die Mutations-Wahrscheinlichkeit, desto häufiger werden Mutationen eingesetzt. Es kann gezeigt werden, dass durch Reproduktion die Fitness kontinuierlich verbessert, durch Ausführung einer Mutation oft verschlechtert wird.
Ziel der Mutation ist, den kontinuierlichen Fortschritt des Individuums zum nächsten Optimum zu stören, um von veränderter Position die Verbesserung neu zu starten.

Stop-Kriterium
Als Stop-Kriterium kann das Erreichen eines asymptotischen Schwellwerts dienen, in den meisten Fällen wird eine zu durchlaufende Zyklenzahl vorgegeben.

Vor-, Nachteile
Der trial-and-error-Ansatz ist numerisch nicht komplex und vielseitig einsetzbar.

Die Effizienz von GA hängt wesentlich von der gewählten Variante für Kreuzung, Mutation und Reproduktion ab, aber auch von einer guten Software-Implementierung.

Die Programmierung der zentralen GA-Routinen ist in der Regel problemlos auszuführen, es ist jedoch nicht möglich, ein Programm für diverse Anwendungsfälle zu erstellen.
Für den aktuellen Fall muss jeweils neu Transformation und Binärcodierung der Variablen x_i und die Berechnung der Fitness explizit implementiert werden.

Bei Optimierung wird schnell das globale Optimum angezeigt, die exakte Lage wird, ähnlich der Simplex-Optimierung, nicht erhalten. Diese muss in einem weiteren Ansatz z.B. mit Response-Surface-Modelling in engerem Untersuchungsbereich ermittelt werden.

Anwendungsbeispiele

Der GA-Einsatz ist in diversen wirtschaftswissenschaftlichen oder technischen Bereichen häufig beschrieben. Insbesondere zur Optimierung technischer Prozesse wird GA auch in der Analytischen Chemie oft verwendet.

- Dies kann z.B. die Festlegung der Veraschungsbedingungen der elektrothermischen Atomabsorption (AAS) sein. Diese hängt von den beiden Faktoren Veraschungs-Temperatur [200, 1000 °C] und -Zeit [5, 50 s] ab.
Die GA-Software legt entsprechend den beiden Chromosomen des aktuellen Individuums die Faktoren vor. Der Anwender führt die dazu gehörige AAS-Messung durch und gibt interaktiv die Fitness (z.B. Absorption / (1+Untergrund)) in die GA-Software ein.

- Häufig beschrieben ist die Bestimmung der optimalen Wellenlängen in der multivariaten Kalibration z.B. aus UV- oder IR-Spektren [3, 4].
Sind von 10 registrierten UV-Wellenlängen die 3 besten für 3 Komponenten zu ermitteln, wird ein Chromosom mit 10 Genen eingesetzt, z.B. für ein aktuelles Individuum 0100010100, d.h. mit der Wahl der 2., 6. und 8. Wellenlänge. Als Fitness dient ein Maß für die Güte der multivariaten Kalibration. Hierzu kann z.B. der RSS-Wert verwendet werden, d.h. die quadratische Abweichung der realen Kalibrator-Konzentration von der geschätzten Konzentration z.B. nach multipler, linearer Regression [5]. Es gilt

$$RSS = \sum_{i=1}^{n} \left(c_{ik} - \hat{c}_{ik} \right)^2$$ (bzw. RSS-Summe über alle Komponenten)

mit c_{ik}: Kalibrator-Konzentration der Komponente k, $\hat{c}_{ij}$: geschätzte Kalibrator-Konzentration durch das Regressions-Modell, n: Zahl der Kalibrationsproben

Im obigen Fall steht beim aktuellen Individuum die Wahl des Absorptions-Signale der 2., 6., 8. registrierten Wellenlänge. Mit diesen wird die multiple, lineare Regression aufgebaut und daraus die aktuellen, resultierenden Kalibrator-Konzentrationen geschätzt. Je besser die gewählten Wellenlängen für die Kalibration geeignet sind, desto kleiner ist die Abweichung von realer und geschätzter Konzentration d.h. der resultierende RSS-Wert.

Literatur

[1] Goldberg, D.E., Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA, 1989
[2] Gerdes, I., Klawonn, F., Kruse, R., Evolutionäre Algorithmen: Computational Intelligence, Vieweg+Teubner Verlag, Wiesbaden, 2004
[3] Khalid A.M. Attia, Mohammed W.I. Nassar, Mohamed B. El-Zeiny, Ahmed Serag, Firefly algorithm versus genetic algorithm as powerful variable selection tools and their effect on different multivariate calibration models in spectroscopy: A comparative study, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 170, 117-123, 2017
[4] Zhao, Y., Wang, S.-H., Li, Z., Cao, F.-Y., Pei, Z.-Y., A Novel Interval Integer Genetic Algorithm Used for Simultaneously Selecting Wavelengths and Pre-processing Methods, Chinese Journal of Analytical Chemistry, 44/9, e1609-e1616, 2016
[5] Reh, E., Chemometrie: Grundlagen der Statistik, Numerischen Mathematik und Software-Anwendung in der Chemie, de Gruyter, Berlin, 2017