Die Höhe von binären Suchbäumen

Ausarbeitung zum Seminar zu Stochastischen Rekursionsgleichungen im WS 2011/2012

Sandra Uhlenbrock

03.11.2011

Die folgende Ausarbeitung wird, basierend auf „Branching Processes and Their Applications in the Analysis of Tree Structures and Tree Algorithms“ von Luc Devroye, mit Hilfe von Verzweigungsprozessen ein Resultat über die Höhe von binären Suchbäumen erläutern.

Galton-Watson Prozesse

Wir betrachten ein Modell für das Wachstum einer Population. Beginnend mit einem Urahn in Generation 0 bezeichne Z_n für $n \geq 0$ die Anzahl an Individuen in der n-ten Generation, wobei für die Reproduktion folgende Eigenschaften gelten:

(i) Jedes Individuum überlebt genau eine Zeiteinheit und produziert am Ende dieser Zeiteinheit eine zufällige Anzahl von Nachkommen.

(ii) Die Anzahl der Nachkommen ist für jedes Individuum identisch verteilt. Sei Z eine Zufallsgröße, die gemäß dieser Verteilung verteilt ist, und setze $p_i = P(Z = i)$.

(iii) Die Nachkommen reproduzieren unabhängig voneinander.

Dann heißt $(Z_n)_{n\in\mathbb{N}_0}$ *Galton-Watson Prozess.*

Satz 1. *In einem Galton-Watson Prozess mit $E(Z) > 1$ gilt für die Aussterbewahrscheinlichkeit*

$$P(Z_n = 0 \text{ für ein } n) = \lim_{n\to\infty} P(Z_n = 0) < 1.$$

Falls $E(Z) \leq 1$, gilt $P(Z_n = 0$ für ein $n) = 1$, außer im entarteten Fall $p_1 = 1$ (d.h. jede Generation besteht aus genau einem Individuum).

Binäre Suchbäume

Die Funktionsweise eines binären Suchbaums wird durch das folgende Beispiel deutlich.

Beispiel. Wir geben die Zahlen von 0 bis 9 in der folgenden Reihenfolge vor: 5 2 6 8 1 4 3 9 7 0. Dies ergibt den in Abbildung 1 dargestellten binären Suchbaum.

Wähle gemäß der Gleichverteilung eine zufällige Permutation von $\{1, \ldots, n\}$ und konstruiere den dazu gehörigen binären Suchbaum. Dieser heißt dann *zufälliger binärer Suchbaum.* H_n bezeichne den Abstand von der Wurzel zum am weitesten entfernten Knoten des Baumes.
Ziel dieser Ausarbeitung ist es, den folgenden Satz zu zeigen.

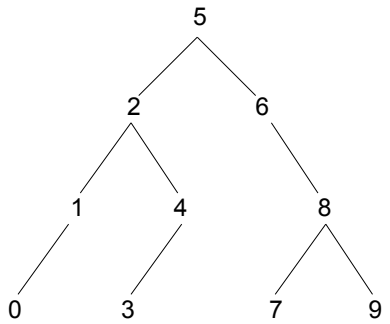

Abbildung 1: Binärer Suchbaum

Satz 2. ***(Devroye, 1986, 1987)***
In einem zufälligen binären Suchbaum T mit n Knoten gilt

$$\frac{H_n}{\log n} \xrightarrow{P} \gamma = 4,31107...$$

Beweis. Zunächst führen wir eine neue Darstellung eines zufälligen binären Suchbaums T ein, mit der wir von den konkreten Zahlen in T abstrahieren, um das Augenmerk auf die Struktur des Baums zu legen. Aus dieser Darstellung gewinnen wir dann eine nützliche Charakterisierung, wann $H_n \geq k$ für ein k gilt.
Um die Darstellung zu erhalten, markieren wir jeden Knoten i in T mit der Größe V_i des an diesem Knoten beginnenden Teilbaums. Diesen Baum nennen wir T'.
Die Wurzel von T' ist nun mit n markiert. Da der Wert der Wurzel gleichverteilt auf $\{1, \ldots, n\}$ ist, ist die Anzahl N der Knoten im linken Teilbaum gleichverteilt auf $\{0, \ldots, n-1\}$. Das heißt $N \stackrel{d}{=} \lfloor nU \rfloor$, wobei U gleichverteilt auf $[0, 1]$ ist. Für den rechten Teilbaum folgt analog $\lfloor n(1-U) \rfloor$.
Wenn man dies auf die anderen Knoten fortsetzt, erhält man Zufallsgrößen $U_1, U_2, \ldots$, welche die Gestalt von T festlegen. Außerdem ist die Markierung V_i eines beliebigen Knotens i mit Abstand k von der Wurzel in T' genauso verteilt wie $\lfloor \ldots \lfloor \lfloor nU_1 \rfloor U_2 \rfloor \ldots U_k \rfloor$.
Mit dieser Darstellung lässt sich die Höhe eines solchen Baumes besser analysieren.
Sei H_n die Höhe von T, wobei $|T| = n$. Dann gilt $H_n \geq k$ genau dann, wenn eine der 2^k Markierungen V_i der Knoten mit Abstand k von der Wurzel einen Wert größer oder gleich 1 hat, d.h.

$$H_n \geq k \Leftrightarrow \max_{1 \leq i \leq 2^k} V_i \geq 1.$$

Um die Behauptung zu zeigen, werden wir nun eine obere und eine untere Schranke für H_n angeben. Für die obere Schranke, werden wir die soeben gewonnene Charakterisierung benutzen. Im Beweis der unteren Schranke werden Galton-Watson Prozesse eine zentrale Rolle spielen.

Die obere Schranke

Es gilt

$$\begin{aligned}
P(H_n \geq k) &= P(\max_{1\leq i\leq 2^k} V_i \geq 1) \\
&= P(V_i \geq 1 \text{ für ein } 1 \leq i \leq 2^k) \\
&\leq \sum_{i=1}^{2^k} P(V_i \geq 1) = 2^k \cdot P(V_1 \geq 1) \\
&= 2^k \cdot P(\lfloor \ldots \lfloor \lfloor nU_1 \rfloor U_2 \rfloor \ldots U_k \rfloor \geq 1) \\
&\leq 2^k \cdot P(n \cdot \prod_{i=1}^{k} U_i \geq 1) \\
&= 2^k \cdot P(n \cdot e^{-G_k} \geq 1), \text{ wobei } G_k \; Gamma(k,1)\text{-verteilt ist} \\
&= 2^k \cdot P(G_k \leq \log n).
\end{aligned}$$

Ziel ist nun, das kleinste k zu finden, so dass die obere Schranke für $n \to \infty$ gegen 0 konvergiert, da dann auch $P(H_n \geq k)$ für $n \to \infty$ gegen 0 konvergiert.
Für $k = \log n$ ist die obere Schranke von der Ordnung $2^{\log n}$, da $P(G_{\log n} \leq \log n)$
$= P(G_{\log n} \leq E(G_{\log n})) \nrightarrow 0$. Das heißt für $k = \log n$ konvergiert die obere Schranke nicht gegen 0. Wir testen also, was für $k \sim c \cdot \log n$ für ein $c > 1$ passiert. Die Arbeit besteht nun darin, die linke Tailwahrscheinlichkeit der Gammaverteilung abzuschätzen und damit zu zeigen, dass die obere Schranke gegen 0 konvergiert. Dazu benutzen wir die folgende Ungleichung

$$1 \leq \frac{P(G_k \leq y)}{\frac{y^k e^{-y}}{k!}} \leq \frac{1}{1 - \frac{y}{k+1}},$$

wobei die untere Schranke für alle $y > 0$ und die obere Schranke für $0 < y < k+1$ gilt. Hieraus folgt insbesondere

$$P(G_k \leq \log n) \leq \frac{(\log n)^k}{n \cdot k!} \cdot \frac{1}{1 - \frac{\log n}{k+1}}$$

für $\log n < k + 1$. Wähle nun $k = \lceil c \cdot \log n \rceil$ für ein $c > 1$, dann folgt

$$\begin{aligned}
P(H_n \geq k) &= 2^k \cdot P(G_k \leq \log n) \\
&\leq \frac{(2 \cdot \log n)^k}{n \cdot k!} \cdot \frac{1 + o(1)}{1 - \frac{1}{c}} \\
&\leq n^{-1} \cdot \frac{(2 \cdot \log n)^k}{(\frac{k}{e})^k} \cdot \frac{1 + o(1)}{1 - \frac{1}{c}} \\
&= n^{-1} \cdot \left(2e \frac{\log n}{k}\right)^k \cdot \frac{1 + o(1)}{1 - \frac{1}{c}} \\
&\leq n^{-1} \left(\frac{2e}{c}\right)^{c \cdot \log n} \cdot \frac{1 + o(1)}{1 - \frac{1}{c}} \\
&= \left(\frac{1}{e}\left(\frac{2e}{c}\right)^c\right)^{\log n} \cdot \frac{1 + o(1)}{1 - \frac{1}{c}} \\
&\to 0 \text{ , falls } \frac{1}{e}\left(\frac{2e}{c}\right)^c < 1.
\end{aligned}$$

Sei $\gamma = 4,31107...$ die einzige Lösung größer als 1 von $\frac{1}{e}\left(\frac{2e}{c}\right)^c = 1$. Dann folgt
$\lim_{n\to\infty} P(H_n > c \cdot \log n) = 0$ für alle $c > \gamma$. Eine Verfeinerung des Arguments mit präziserer

Benutzung der Stirling Ungleichung zeigt auch, dass

$$\lim_{n\to\infty} P(H_n > \gamma \cdot \log n) = 0.$$

Die untere Schranke

Wir zeigen nun, dass $\lim_{n\to\infty} P(H_n \geq (\gamma - \epsilon)\log n) = 1$ für alle $\epsilon > 0$.
Sei also $\epsilon > 0$. Um diese obere Schranke zu zeigen, müssen wir im markierten Baum einen Pfad finden, der bei Abstand $k = \lfloor(\gamma - \epsilon)\log n\rfloor$ von der Wurzel eine Markierung mit Wert größer oder gleich 1 hat.
Es funktioniert allerdings nicht, wenn man dem Pfad folgt, der dadurch gegeben ist, dass man bei jedem Knoten in den größeren Teilbaum hinein geht, da dieser Pfad nur von der Größenordnung $c\log n$ für $c \approx 3,25$ ist.
Wir werden für diesen Teil des Beweises Verzweigungsprozesse benutzen. Dazu werden wir einen Galton-Watson Prozess definieren, dessen Baum einer „ausgedünnten“ Version von T entspricht und zeigen, dass dieser Prozess mit positiver Wahrscheinlichkeit überlebt.

Die Wurzel von T sei der Urahn im Galton-Watson Prozess. Alle Kinder in T, die l Level von der Wurzel entfernt sind, werden die direkten Nachkommen der Wurzel, falls das Produkt ihrer „Teilungsvariablen“ U_i größer oder gleich d^l ist, für eine gegebene Konstante d. Das Reproduktionsverhalten der anderen Knoten ergibt sind analog, da die U_i unabhängig und identisch verteilt sind. Das heißt, die Anzahl der Nachkommen ist binominalverteilt mit $n = 2^l$ und $p = P(U_1 \dots U_l > d^l)$. Falls T unendlich ist, würde der zugehörige Galton-Watson Prozess mit Wahrscheinlichkeit $1-q > 0$ für ein $q > 0$ überleben, falls das Reproduktionsmittel, also die erwartete Anzahl von Nachkommen pro Knoten, größer als 1 ist. Dies ist hier

$$\begin{aligned}
2^l P(U_1 \dots U_l > d^l) &= 2^l P(G_l < l \cdot \log\left(\frac{1}{d}\right)), \text{ wobei } G_l \sim Gamma(l,1) \\
&\geq \frac{\left(2ld\log\left(\frac{1}{d}\right)\right)^l}{l!} \\
&\sim \frac{\left(2ed\log\left(\frac{1}{d}\right)\right)^l}{\sqrt{2\pi l}} \\
&> 1 \text{ für } l \text{ groß genug, falls } 2ed\log\left(\frac{1}{d}\right) > 1.
\end{aligned}$$

Dies ist erfüllt, wenn wir $d = e^{-\frac{1}{c}}$ für $1 < c < \gamma$ wählen.
Das heißt mit Wahrscheinlichkeit größer oder gleich $1-q > 0$ existiert in T ein Knoten mit Abstand kl von der Wurzel mit Wert $V = \lfloor \dots \lfloor\lfloor nU_1\rfloor U_2\rfloor \dots U_{kl}\rfloor \geq nU_1U_2\dots U_{kl} - kl \geq nd^{kl} - kl = ne^{-\frac{kl}{c}} - kl$, da man beim Abrunden in jedem Schritt maximal eine Einheit verlieren kann.
Das heißt $P(H_n \geq kl) \geq 1 - q$, falls $ne^{-\frac{kl}{c}} - kl \geq 1$. Dies ist zum Beispiel für $kl = c'\log n - \theta l$, $c' < c, \theta \in [0,1)$ für alle genügend großen n erfüllt. Da c' beliebig nah an c liegt und c beliebig nah an γ ist, folgt

$$\liminf_{n\to\infty} P(H_n > (\gamma - \epsilon)\log n) \geq 1 - q \text{ für alle } \epsilon > 0 \text{ und ein } q < 1.$$

Dies zeigt jedoch noch nicht die Behauptung, da q beliebig nah an 1 liegen kann. Mit etwas zusätzlichem Aufwand kann man jedoch auch zeigen, dass $lim_{n\to\infty} P(H_n \geq (\gamma - \epsilon)\log n) = 1$. Dies geschieht dadurch, dass man geschickt zwei Ereignisse A und B wählt, sodass, falls A und B beide eintreten, ein Knoten mit Wert größer oder gleich 1 mit genügend großem Abstand von der Wurzel existiert, und sodass die Wahrscheinlichkeiten, dass diese Ereignisse nicht eintreten, jeweils nahe bei 0 liegen.
Aus der oberen und unteren Schranke folgt nun, dass

$$\frac{H_n}{\log n} \xrightarrow{P} \gamma.$$

□

Quadtrees

Das obige Resultat lässt sich auf Quadtrees verallgemeinern.
Ein Quadtree im $\mathbb{R}^d$ zeichnet sich dadurch aus, dass jedes Datum, als Knoten im Baum, 2^d Teilbäume hat. Das Einfügen funktioniert genauso, wie beim binären Suchbaum.

Beispiel. Ein 2-dimensionaler Quadtree mit den Eingaben $(1, 1)$ $(0, -1)$ $(-1, 1.5)$ ist in Abbildung 2 grafisch dargestellt.

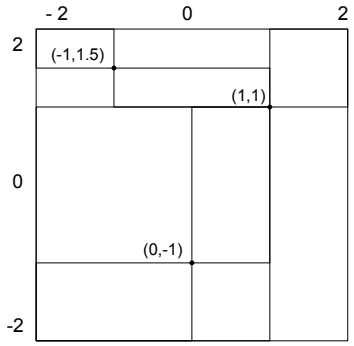

Abbildung 2: Quadtree

Wenn man annimmt, dass der Quadtree auf Basis von unabhängig identisch gleichmäßig verteilten Zufallsgrößen konstruiert wurde, kann man wieder zeigen, dass

$$\frac{H_n}{\log n} \xrightarrow{P} \frac{\gamma}{d}.$$