On the Uniform Distribution of Strings

Sébastien Rebecchi and Jean-Michel Jolion

PSC 2008, Prague, Czech Republic
September 2, 2008
Introduction

How to describe the data?

- Structures:
  + representational capabilities,
  - lack of mathematical tools;

- feature vectors:
  + powerful statistical algorithms,
  - representational capabilities;

⇒ reconcile the two approaches;
⇒ need to define a statistical characterization of spaces of structures.

We introduce the uniform distribution of strings.
Notations

- $A$ alphabet;
- $|A|$ cardinal of $A$. 
Notations

- $|X|$ length of the string $X$ over $A$;
- $A^n$ set of strings of length $n$ over $A$;
- $A^{\leq n}$ set of strings of length at most $n$ over $A$;
- $X_i$ $i$-th letter of $X$. 
Uniform distribution of strings

First approach: equiprobability.

- $U$ over $A^n = \text{concatenation of } n \ U \over A$
  \[ P(X) = |A|^{-n}; \]
- generation in $O(n)$;
- probability in $O(1)$;
- preservation under concatenation:
  \[ (X \sim U \over A^n) \land (I \sim U \over A) \implies Xl \sim U \over A^{n+1}. \]
Uniform distribution of strings

Second approach: normalized measure.

• Let $S$ be a set;
• $E \subseteq S$:
  
  $P(E) = \frac{\mu(E)}{\mu(S)}$;

• examples:
  • $S \subseteq \mathbb{N}$, $S$ finite, $\mu = \text{cardinality}$,
  • $S \subseteq \mathbb{R}$, $S$ bounded, $\mu = \text{Lebesgue measure}$.
σ-algebra

σ-algebra over $S = \text{set of subsets of } S$ that is

- non empty;
- closed under complements;
- closed under countable unions.

If $S$ is countable, then $\text{powerset}(S)$ is the only σ-algebra over $S$ containing all singletons $\{x\}, x \in S$. 
Uniform distribution of strings

Measure

Measure $\mu$ over $\sigma = \text{function } \sigma \rightarrow \mathbb{R}^+ \cup \{\infty\}$ that is

- 0 for $\{\}$;
- additive under countable disjoint unions.

$\mu(\{x\}) = \text{notation } \mu(x)$, $\{x\} \in \sigma$. 
Uniform distribution w.r.t. $\mu$: $\forall E \in \sigma$:

$$P(E) = \frac{\mu(E)}{\mu(S)}.$$ 

$P(\{x\}) = \text{notation } P(x), \{x\} \in \sigma.$
String measure

- \( \lambda \notin A \) denotes the empty letter;
- we assume the measure \( \mu_A \) over \( \text{powerset}(A \cup \{\lambda\}) \);
- for \( n \in \mathbb{N} \), we define the measure \( \mu^n \) over \( \text{powerset}(A^{\leq n}) \).
String measure

- String of length at most $n$ over $A = \text{canonical representation of a set of sequences composed of } n \text{ elements of } A \cup \{\lambda\}$;

- example:
  - $A = \{a, b\}$,
  - $n = 3$,
  - $ab''='' \{\lambda ab, a\lambda b, ab\lambda\}$. 
String measure

\(\forall X \in A^{\leq n}:
\)

\[
\mu^n(X) = \binom{n}{|X|} \times \prod_{i=1}^{|X|} \mu_A(X_i) \times \mu_A(\lambda)^{n-|X|}.
\]
Total measure:

$$\mu^n(A^{\leq n}) = \mu_A(A \cup \{\lambda\})^n.$$ 

⇒ Probability of a string in $O(n)$. 

S. Rebecchi and J.-M. Jolion (LIRIS) On the Uniform Distribution of Strings PSC 2008 13 / 17
Preservation, generation

Preservation under concatenation:

\[(X \sim U \text{ w.r.t. } \mu^n) \land (l \sim U \text{ w.r.t. } \mu_A) \implies Xl \sim U \text{ w.r.t. } \mu^{n+1}.\]

\[\implies \text{Generation of a string in } O(n).\]
Generation

**Input:** $n \in \mathbb{N}$.

**Output:** A string uniform w.r.t. $\mu^n$.

begin
\begin{align*}
P_A &\leftarrow \text{uniform distribution w.r.t. } \mu_A: \forall l \in A \cup \{\lambda\}: \\
P_A(l) &= \frac{\mu_A(l)}{\mu_A(A \cup \{\lambda\})};
\end{align*}

$X \leftarrow$ empty string;

for $i \leftarrow 1 \text{ à } n$ do
\begin{itemize}
  \item $l \leftarrow \text{random choice according to } P_A$;
  \item $X \leftarrow Xl$;
\end{itemize}
end

return $X$;
end
Uniform distribution of strings

Unification

First approach = second approach with:

- \( \mu_A(\lambda) = 0 \) \( \iff \) \( P^n(A \leq n-1) = 0 \) if \( n > 0 \);
- \( \mu_A(l) = \mu_A(m) \), \( \forall l, m \in A \).
Conclusion

- Uniform string = concatenation of uniform letters;
- simple but relevant measure;
- easy to extend to ordered trees;
- statistical test;
- how to sum for CLT?