Proof-Pattern Recognition and Lemma Discovery in ACL2

Jónathan Heras
(joint work with K. Komendantskaya, M. Johansson and E. Maclean)

University of Dundee
http://staff.computing.dundee.ac.uk/jheras/acl2ml/

31 January 2014
Scottish Theorem Proving seminar

Outline

Outline

ACL2

ACL2 (A Computational Logic for an Applicative Common Lisp) is the successor of the Boyer-Moore theorem prover.

ACL2

ACL2 (A Computational Logic for an Applicative Common Lisp) is the successor of the Boyer-Moore theorem prover.
ACL2 is . . .

- . . . a programming language (an extension of an applicative subset of Common Lisp).
- . . . a logic (an untyped first-order logic with equality).
- . . . a theorem prover.

ACL2

ACL2 (A Computational Logic for an Applicative Common Lisp) is the successor of the Boyer-Moore theorem prover.
ACL2 is ...

- ... a programming language (an extension of an applicative subset of Common Lisp).
- ... a logic (an untyped first-order logic with equality).
- ... a theorem prover.

Applications of ACL2:

- Software and Hardware Verification (microprocessors, flash memories, JVM-like bytecode, ...)

Proving in ACL2

ACL2 has features of both interactive and automated theorem provers.

Proving in ACL2

ACL2 has features of both interactive and automated theorem provers.

- Automatic: once a proof attempt is started, the user can no longer interact with ACL2.

Proving in ACL2

ACL2 has features of both interactive and automated theorem provers.

- Automatic: once a proof attempt is started, the user can no longer interact with ACL2.
- Interactive: the user has to supply a suitable collection of definitions and auxiliary lemmas to guide ACL2.

Challenges

- . . . size of ACL2 library stands on the way of efficient knowledge reuse;
- . . . manual handling of proofs, strategies, libraries becomes difficult;
- . . . team-development is hard;
- . . . comparison of proof similarities is hard;
- . . . discovery of auxiliary lemmas can be difficult.

Challenges

- ...size of ACL2 library stands on the way of efficient knowledge reuse;
- ...manual handling of proofs, strategies, libraries becomes difficult;
- ...team-development is hard;
- ...comparison of proof similarities is hard;
- ...discovery of auxiliary lemmas can be difficult.

What can we do?

- Statistical methods can discover patterns in proofs but are weak for conceptualisation.
- Symbolic methods (Proof planning, lemma discovery) can conceptualise but have limitations.

Challenges

- ...size of ACL2 library stands on the way of efficient knowledge reuse;
- ...manual handling of proofs, strategies, libraries becomes difficult;
- ...team-development is hard;
- ...comparison of proof similarities is hard;
- ...discovery of auxiliary lemmas can be difficult.

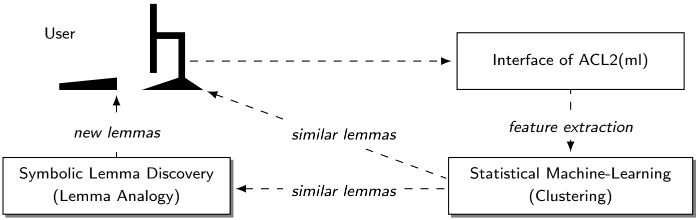
What can we do?

- Statistical methods can discover patterns in proofs but are weak for conceptualisation.
- Symbolic methods (Proof planning, lemma discovery) can conceptualise but have limitations.
- Combination of statistical and symbolic methods:
  - Statistical methods can take advantage of symbolic methods to conceptualise results.
  - Symbolic tools can use statistical results for efficient lemma discovery.

Outline

ACL2(ml)

. . . in [2013, Proceedings of LPAR'13]



F.1. works on the background of Emacs extracting some low-level features from ACL2 definitions and theorems.

F.2. automatically sends the gathered statistics to a chosen machine-learning interface and triggers execution of a clustering algorithm of user's choice;

F.3. does some post-processing of the results and

- **F.3.a** displays families of related proofs (or definitions) to the user.
- **F.3.b** uses the families of related proofs to discover new lemmas.

Outline

Extracting features from ACL2

1. Feature extraction:

Extracting features from ACL2

1. Feature extraction:
   - We extract features directly from term trees of ACL2 terms.

Definition (Term tree)

A variable or a constant is represented by a tree consisting of one single node, labelled by the variable or the constant itself. A function application $f(t_1, \ldots, t_n)$ is represented by the tree with the root node labelled by f, and its immediate subtrees given by trees representing $t_1, \ldots, t_n$.

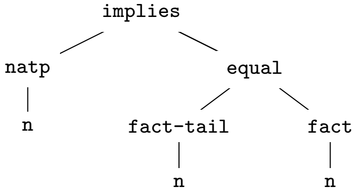
Extracting features from ACL2

1. Feature extraction:
   - We extract features directly from term trees of ACL2 terms.

Definition (Term tree)

A variable or a constant is represented by a tree consisting of one single node, labelled by the variable or the constant itself. A function application $f(t_1, \ldots, t_n)$ is represented by the tree with the root node labelled by f, and its immediate subtrees given by trees representing $t_1, \ldots, t_n$.

```
(implies (natp n) (equal (fact-tail n) (fact n)))
```



ACL2(ml) term tree matrices

We have devised a compact feature extraction method.

ACL2(ml) term tree matrices

We have devised a compact feature extraction method.

Definition (Term tree depth level)

Given a term tree T, the *depth* of the node t in T, denoted by *depth(t)*, is defined as follows:
$-$ $depth(t) = 0$, if t is a root node;
$-$ $depth(t) = n + 1$, where n is the depth of the parent node of t.

ACL2(ml) term tree matrices

We have devised a compact feature extraction method.

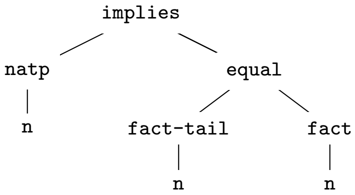
Definition (Term tree depth level)

Given a term tree T, the *depth* of the node t in T, denoted by *depth(t)*, is defined as follows:
− $depth(t) = 0$, if t is a root node;
− $depth(t) = n + 1$, where n is the depth of the parent node of t.

Definition (ACL2(ml) term tree matrices)

Given a term tree T for a term with signature Σ, and a function $[.] : \Sigma \rightarrow \mathbb{Q}$, the ACL2(ml) term tree matrix M_T is a 7×7 matrix that satisfies the following conditions:
− the $(0, j)$-th entry of M_T is a number $[t]$, such that t is a node in T, t is a variable and $depth(t) = j$.
− the (i, j)-th entry of M_T $(i \neq 0)$ is a number $[t]$, such that t is a node in T, t has arity $i + 1$ and $depth(t) = j$.

An example



	variables	arity 0	arity 1	arity 2
*td*0	0	0	0	[implies]
*td*1	0	0	[natp]	[equal]
*td*2	[n]	0	[fact-tail]::[fact]	0
*td*3	[n]::[n]	0	0	0

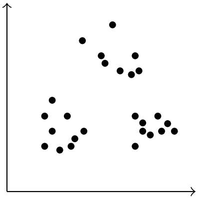
Clustering in ACL2(ml)

We have integrated Emacs with a variety of clustering algorithms:

Clustering in ACL2(ml)

We have integrated Emacs with a variety of clustering algorithms:

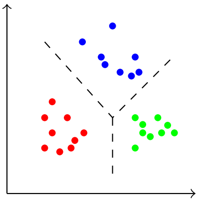
- Unsupervised machine learning technique:



Clustering in ACL2(ml)

We have integrated Emacs with a variety of clustering algorithms:

- Unsupervised machine learning technique:

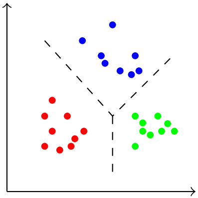


- Engines: Matlab, Weka, Octave, R, ...

Clustering in ACL2(ml)

We have integrated Emacs with a variety of clustering algorithms:

- Unsupervised machine learning technique:

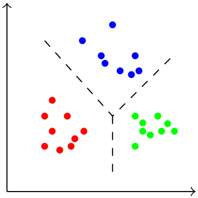


- Engines: Matlab, Weka, Octave, R, ...

Clustering in ACL2(ml)

We have integrated Emacs with a variety of clustering algorithms:

- Unsupervised machine learning technique:

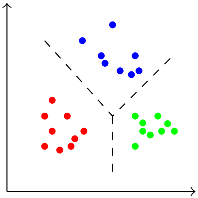


- Engines: Matlab, Weka, Octave, R, . . .
- Algorithms: K-means, simple Expectation Maximisation, . . .

Clustering in ACL2(ml)

We have integrated Emacs with a variety of clustering algorithms:

- Unsupervised machine learning technique:



- Engines: Matlab, Weka, Octave, R, ...
- Algorithms: K-means, simple Expectation Maximisation, ...

Recurrent clustering

How is the function [.] defined?

Definition (Function [.])

Given the nth term definition of the library (call the term t), a function [.] is inductively defined for every symbol s in t as follows:
– $[s] = i$, if s is the ith distinct variable in t (formulas are implicitly universally quantified);
– $[s] = -[m]$, if t is a recursive definition defining the function s with measure function m;
– $[s] = k$, if s is a function imported from CLISP; and $[s] = k$ in the figure below;
– $[s] = 5 + 2 \times j + p$, where C_j is a cluster obtained as a result of definition clustering with granularity 3 for library definitions 1 to $n - 1$, $s \in C_j$ and p is the proximity value of s in C_j.

$*$ Type recognisers ($r = \{$symbolp, characterp, stringp, consp, acl2-numberp, integerp, rationalp, complex-rationalp$\}$): $[r_i] = 1 + \sum_{j=1}^{i} \frac{1}{10 \times 2^{j-1}}$ (where r_i is the ith element of r).

$*$ Constructors ($c = \{$cons, complex$\}$): $[c_i] = 2 + \sum_{j=1}^{i} \frac{1}{10 \times 2^{j-1}}$.

$*$ Accessors ($a^1 = \{$car, cdr$\}$, $a^2 = \{$denominator, numerator$\}$, $a^3 = \{$realpart, imagpart$\}$): $[a^j_i] = 3 + \frac{1}{10 \times j} + \frac{i-1}{100}$.

$*$ Operations on numbers ($o = \{$unary-/, unary–, binary-+, binary-*$\}$): $[o_i] = 4 + \sum_{j=1}^{i} \frac{1}{10 \times 2^{j-1}}$.

$*$ Integers and rational numbers: $[0] = 4.3$, $[n] = 4.3 + \frac{|n|}{10}$ (with $n \neq 0$ and $|n| < 1$) and $[n] = 4.3 + \frac{1}{100 * |n|}$ (with $n \neq 0$ and $|n| \geq 1$).

$*$ Boolean operations ($b = \{$equal, if, ¡$\}$): $[b_i] = 5 + \sum_{j=1}^{i} \frac{1}{10 \times 2^{j-1}}$.

Demo

- Finding similar theorems across libraries.
- Obtaining more precise clusters.
- Finding similar definitions across libraries.

Outline

Lemma analogy in ACL2(ml)

Can we use the output of the statistical side of ACL2(ml) to generate useful lemmas?

Lemma analogy in ACL2(ml)

Can we use the output of the statistical side of ACL2(ml) to generate useful lemmas?

Terminology:

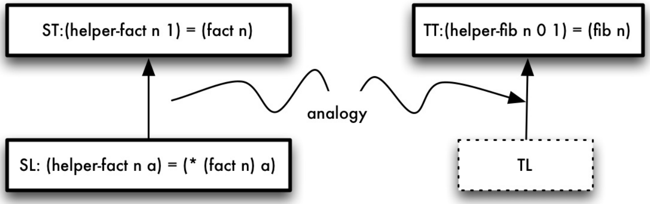
- Target Theorem (TT): the theorem that we want to prove.
- Source Theorem (ST): theorem suggested as similar to TT.
- Source Lemma (SL): a user-supplied lemma to prove ST.

Lemma analogy in ACL2(ml)

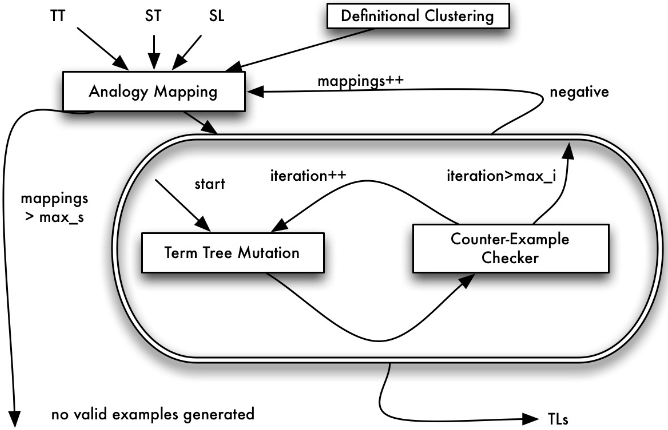
Can we use the output of the statistical side of ACL2(ml) to generate useful lemmas?

Terminology:

- Target Theorem (TT): the theorem that we want to prove.
- Source Theorem (ST): theorem suggested as similar to TT.
- Source Lemma (SL): a user-supplied lemma to prove ST.



Overview of the process



Analogy mapping

Definition (Analogy Mapping $\mathcal{A}$)

For all symbols $s_1, \ldots, s_n$ occurring in the current ST, the set of admissible analogy mappings is the set of all mappings $\mathcal{A}$ such that
- $\mathcal{A}(s_i) = s_i$ for all shared background symbols; otherwise:
- $\mathcal{A}(s_i) = s_j$ for all combinations of $i, j \in 1 \ldots n$, such that s_i and s_j belong to the same cluster in the last iteration of definition clustering.

Analogy mapping

Definition (Analogy Mapping $\mathcal{A}$)

For all symbols $s_1, \ldots, s_n$ occurring in the current ST, the set of admissible analogy mappings is the set of all mappings $\mathcal{A}$ such that
- $\mathcal{A}(s_i) = s_i$ for all shared background symbols; otherwise:
- $\mathcal{A}(s_i) = s_j$ for all combinations of $i, j \in 1 \ldots n$, such that s_i and s_j belong to the same cluster in the last iteration of definition clustering.

Example

For our running example, the shared background theory includes symbols $\{$`+`, `*`,`-`, `1`, `0`$\}$. We thus get a mapping:
$\mathcal{A} = \{\texttt{fact} \mapsto \texttt{fib}, \texttt{helper-fact} \mapsto \texttt{helper-fib}, \texttt{+} \mapsto \texttt{+}, \texttt{1} \mapsto \texttt{1}, \ldots\}$

Term tree mutation

Term tree mutation consists of three iterations:

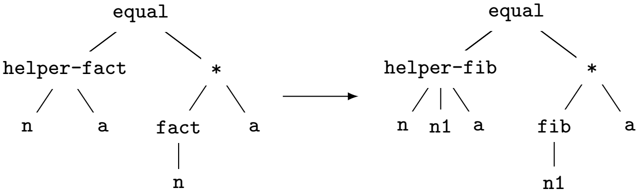
- Tree reconstruction.
- Node expansion.
- Term tree expansion.

Tree reconstruction

Tree Reconstruction phase replaces symbols in the SL with their analogical counterparts.

Tree reconstruction

Tree Reconstruction phase replaces symbols in the SL with their analogical counterparts.

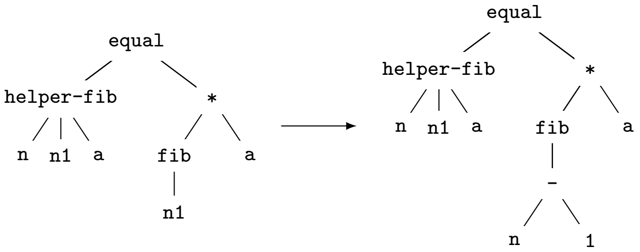


Node expansion

Node expansion phase mutates the term, by synthesising small terms (max depth 2) in place of variables.

Node expansion

Node expansion phase mutates the term, by synthesising small terms (max depth 2) in place of variables.

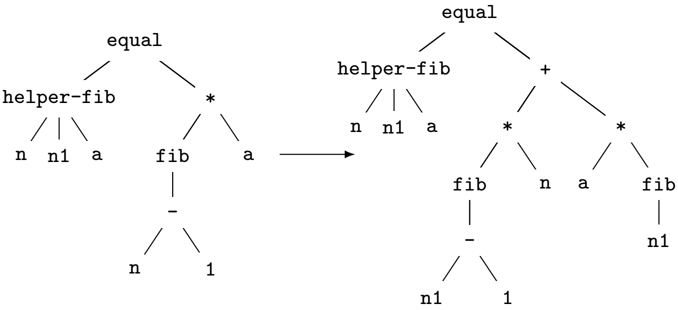


Term Tree Expansion

Term Tree Expansion phase is similar to Node expansion phase, but adding new term structure on the top-level of the term.

Term Tree Expansion

Term Tree Expansion phase is similar to Node expansion phase, but adding new term structure on the top-level of the term.



Using guards to generate preconditions

- Using the lemma analogy tool, ACL2(ml) generates the following suggestion:

```
(equal (helper_fib n j k)
       (+ (* (theta_fib (- n 1)) j) (* (theta_fib n) k))))
```

Using guards to generate preconditions

- Using the lemma analogy tool, ACL2(ml) generates the following suggestion:

```
(equal (helper_fib n j k)
       (+ (* (theta_fib (- n 1)) j) (* (theta_fib n) k))))
```

- This result cannot be directly proven in ACL2, we need some preconditions.

Using guards to generate preconditions

- Using the lemma analogy tool, ACL2(ml) generates the following suggestion:

```
(equal (helper_fib n j k)
       (+ (* (theta_fib (- n 1)) j) (* (theta_fib n) k))))
```

- This result cannot be directly proven in ACL2, we need some preconditions.
- ACL2 is an untyped system, but we can restrict a function to a particular domain using the guard mechanism.

Using guards to generate preconditions

- Using the lemma analogy tool, ACL2(ml) generates the following suggestion:

```
(equal (helper_fib n j k)
       (+ (* (theta_fib (- n 1)) j) (* (theta_fib n) k))))
```

- This result cannot be directly proven in ACL2, we need some preconditions.
- ACL2 is an untyped system, but we can restrict a function to a particular domain using the guard mechanism.
- Guards are optional and several functions do not include them.
- ACL2 recommendation for novices: “novices are often best served by avoiding guards”.

Using guards to generate preconditions

- Using the lemma analogy tool, ACL2(ml) generates the following suggestion:

```
(equal (helper_fib n j k)
       (+ (* (theta_fib (- n 1)) j) (* (theta_fib n) k))))
```

- This result cannot be directly proven in ACL2, we need some preconditions.
- ACL2 is an untyped system, but we can restrict a function to a particular domain using the guard mechanism.
- Guards are optional and several functions do not include them.
- ACL2 recommendation for novices: "novices are often best served by avoiding guards".
- Solution: compute recursively the guards of a function f.

Using guards to generate preconditions

```
(defun helper_fib (n j k)
     (if (zp n) j (if (equal n 1) k (helper_fib (- n 1) k (+ j k)))))

  * zp ->  (natp x)
  * equal -> t
  * + -> (and (acl2-numberp x) (acl2-numberp y))
  * - -> (and (acl2-numberp x) (acl2-numberp y))
```

Using guards to generate preconditions

```
(defun helper_fib (n j k)
     (if (zp n) j (if (equal n 1) k (helper_fib (- n 1) k (+ j k)))))
```

```
  * zp ->  (natp x)
  * equal -> t
  * + -> (and (acl2-numberp x) (acl2-numberp y))
  * - -> (and (acl2-numberp x) (acl2-numberp y))
```

Guards generated for `helper_fib` $\rightarrow$

```
(and (natp n) t (and (acl2-numberp n) (acl2-numberp 1))
     (and (acl2-numberp j) (acl2-numberp k)))
```

$\xrightarrow{simpl}$ `(and (integerp n) (not (< n 0)) (acl2-numberp j) (acl2-numberp k))`

Using guards to generate preconditions

```
(defun helper_fib (n j k)
     (if (zp n) j (if (equal n 1) k (helper_fib (- n 1) k (+ j k)))))

  * zp ->  (natp x)
  * equal -> t
  * + -> (and (acl2-numberp x) (acl2-numberp y))
  * - -> (and (acl2-numberp x) (acl2-numberp y))
```

Guards generated for `helper_fib` $\rightarrow$

```
(and (natp n) t (and (acl2-numberp n) (acl2-numberp 1))
     (and (acl2-numberp j) (acl2-numberp k)))
```

$\xrightarrow{\textit{simpl}}$ `(and (integerp n) (not (< n 0)) (acl2-numberp j) (acl2-numberp k))`

```
(defthm helper_fib_theta_fib
   (equal (helper_fib n j k) (+ (* (theta_fib (- n 1)) j) (* (theta_fib n) k))))
```

Guards:

```
(and (integerp n) (not (< n 0)) (acl2-numberp j) (acl2-numberp k)
     (not (< (+ -1 n) 0)))
```

Demo

- Lemma discovery.
- Guard generation.

Outline

Benefits of ACL2(ml)

- ACL2(ml) statistical and symbolic tools can be switched on/off on user's demand;

Benefits of ACL2(ml)

- ACL2(ml) statistical and symbolic tools can be switched on/off on user's demand;
- ACL2(ml) does not assume any knowledge of machine-learning interfaces from the user;

Benefits of ACL2(ml)

- ACL2(ml) statistical and symbolic tools can be switched on/off on user's demand;
- ACL2(ml) does not assume any knowledge of machine-learning interfaces from the user;
- modular: allows the user to make choices regarding approach to levels of proofs and particular statistical algorithms;

Benefits of ACL2(ml)

- ACL2(ml) statistical and symbolic tools can be switched on/off on user's demand;
- ACL2(ml) does not assume any knowledge of machine-learning interfaces from the user;
- modular: allows the user to make choices regarding approach to levels of proofs and particular statistical algorithms;
- tolerant to mixing and matching different proof libraries and different notation used in proofs across different users.

Benefits of ACL2(ml)

- ACL2(ml) statistical and symbolic tools can be switched on/off on user's demand;
- ACL2(ml) does not assume any knowledge of machine-learning interfaces from the user;
- modular: allows the user to make choices regarding approach to levels of proofs and particular statistical algorithms;
- tolerant to mixing and matching different proof libraries and different notation used in proofs across different users.

Conclusions

- ACL2(ml) combines statistical machine learning (detection of patterns) with symbolic techniques (generation of lemmas).
- ACL2(ml) is different to other tools:
  - its methods of generating the proof-hints interactively and in real-time;
  - its flexible environment for integration of statistical and symbolic techniques.

Further work

- **Different patterns.** Statistical ACL2(ml) groups in the same clusters theorems whose lemmas cannot be mutated to generate any useful lemma.
- **Smaller lemmas.** The lemma analogy tool currently only adds term structure; therefore, it cannot generate smaller lemmas.
- **Conditional lemmas.** Discovering appropriate conditions for generated lemmas is a difficult problem for theory exploration systems.
- **New definitions.** Another big challenge in lemma discovery is invention of new concepts.

Proof-Pattern Recognition and Lemma Discovery in ACL2

Jónathan Heras
(joint work with K. Komendantskaya, M. Johansson and E. Maclean)

University of Dundee
http://staff.computing.dundee.ac.uk/jheras/acl2ml/

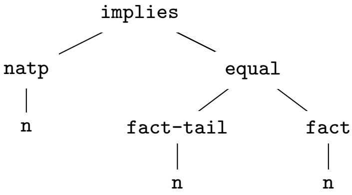
31 January 2014
Scottish Theorem Proving seminar

Methods to represent term-trees

A variety of methods exists: e.g. incidence matrices, adjacency matrices.

Methods to represent term-trees

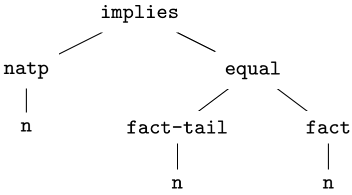
A variety of methods exists: e.g. incidence matrices, adjacency matrices.



	impl	natp	equal	fact-t.	fact	n	n'	n''
impl	0	1	1	0	0	0	0	0
natp	0	0	0	0	0	1	0	0
equal	0	0	0	1	1	0	0	0
fact-t.	0	0	0	0	0	0	1	0
fact	0	0	0	0	0	0	0	1
n	0	0	0	0	0	0	0	0
n'	0	0	0	0	0	0	0	0
n''	0	0	0	0	0	0	0	0

Methods to represent term-trees

A variety of methods exists: e.g. incidence matrices, adjacency matrices.



	impl	natp	equal	fact-t.	fact	n	n'	n''
impl	0	1	1	0	0	0	0	0
natp	0	0	0	0	0	1	0	0
equal	0	0	0	1	1	0	0	0
fact-t.	0	0	0	0	0	0	1	0
fact	0	0	0	0	0	0	0	1
n	0	0	0	0	0	0	0	0
n'	0	0	0	0	0	0	0	0
n''	0	0	0	0	0	0	0	0

Properties of these methods:

- different tree nodes are represented by distinct matrix entries;
- the matrix entries are binary;
- the size of the matrix depend on the tree size; and
- they can grow very large.

Evaluation

Scalability: ACL2(ml) works well with libraries of varied sizes and complexities.

(150 lemmas)	$g = 1$ ($n = 16$)	$g = 2$ ($n = 18$)	$g = 3$ ($n = 21$)	$g = 4$ ($n = 25$)	$g = 5$ ($n = 30$)
fib-fib-tail	$9^{a,b,c}$	$4^{a,b,c}$	$3^{a,c}$	2^{a}	2^{a}

(996 lemmas)	$g = 1$ ($n = 110$)	$g = 2$ ($n = 124$)	$g = 3$ ($n = 142$)	$g = 4$ ($n = 166$)	$g = 5$ ($n = 199$)
fib-fib-tail	$57^{a,b,c}$	$50^{a,b,c}$	$25^{a,b,c}$	$8^{a,b,c}$	$4^{a,b,c}$

Evaluation

Scalability: ACL2(ml) works well with libraries of varied sizes and complexities.

(150 lemmas)	$g = 1$ ($n = 16$)	$g = 2$ ($n = 18$)	$g = 3$ ($n = 21$)	$g = 4$ ($n = 25$)	$g = 5$ ($n = 30$)
fib-fib-tail	$9^{a,b,c}$	$4^{a,b,c}$	$3^{a,c}$	2^{a}	2^{a}

(996 lemmas)	$g = 1$ ($n = 110$)	$g = 2$ ($n = 124$)	$g = 3$ ($n = 142$)	$g = 4$ ($n = 166$)	$g = 5$ ($n = 199$)
fib-fib-tail	$57^{a,b,c}$	$50^{a,b,c}$	$25^{a,b,c}$	$8^{a,b,c}$	$4^{a,b,c}$

Usability of Statistical Suggestions (996 lemmas):

- 37% of clusters can be directly used by the Lemma Analogy tool of ACL2(ml) to mutate lemmas.
- 19% of cluster contain basic theorems whose proofs are similar (based on simplification).
- 15% of clusters contain theorems that use the same lemmas in their proofs.
- 15% of clusters consist of theorems that are used in the proofs of other theorems of the same cluster.
- 14% of clusters do not show a clear correlation that could be reused.

Evaluation II

Modularity: ACL2(ml) provides a flexible environment for integrating statistical and symbolic machine-learning methods.

Evaluation II

Modularity: ACL2(ml) provides a flexible environment for integrating statistical and symbolic machine-learning methods.

Lemma discovery: reduces the combinatorial explosion of theory exploration techniques.

Evaluation II

Modularity: ACL2(ml) provides a flexible environment for integrating statistical and symbolic machine-learning methods.

Lemma discovery: reduces the combinatorial explosion of theory exploration techniques. Comparison with QuickSpec:

		Target						
		fact	*power*	*expt*	*sum*	*sum_sq*	*mult*	*fib*
Source	*fact*	-	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	$\checkmark_2$	$\checkmark_1$
	power	$\checkmark_1$	-	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	$\checkmark_2$	$\checkmark_1$
	expt	$\checkmark_1$	$\checkmark_1$	-	$\checkmark_1$	$\checkmark_1$	$\checkmark_2$	$\checkmark_1$
	sum	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	-	$\checkmark_1$	$\checkmark_2$	$\checkmark_1$
	sum_sq	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	-	$\checkmark_2$	$\checkmark_1$
	mult	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	$\checkmark_1$	-	$\checkmark_1$
	fib	$(\checkmark_2)$	$(\checkmark_2)$	$\times$	$\times$	$\times$	$\times$	-

QuickSpec		
Lemma	Valid	Invalid
$\times$	10	5
$\times$	17	4
$\times$	OoM	OoM
$\times$	7	2
$\times$	7	1
$\times$	200	20
$\times$	OoM	OoM