Carnegie Mellon
DATA PRIVACY LAB

Adding Semantics and Rigor to Association Rule Learning: the GenTree Approach

Yiheng Li and Latanya Sweeney
School of Computer Science, Carnegie Mellon University
http://privacy.cs.cmu.edu/dataprivacy/projects/gentree/

Abstract

Learning "useful" association rules across all attributes of a relational table requires: (1) more rigorous mining than afforded by traditional approaches; and, (2) the invention of knowledge ratings for learned rules, not just statistical ratings. Originally rules were learned over one attribute –e.g., "people who buy diapers tend to buy beer." Then, a hierarchy (whose base values are those originally in the data, and values appearing at higher levels represent increasingly more general concepts) was used to learn generalized association rules –e.g., "people who buy baby products tend to buy controlled substances." Many kinds of rules were still unable to be learned. The work reported herein continues the evolution to its broadest application –learning rules by mixing generalizations across attributes and selecting those whose features convey the maximum expression of information. We term these robust rules –e.g., "people who buy baby products tend to buy controlled substances, use credit cards and make purchases on Saturdays." The most semantically relevant rules are those expressing more depth and/or breadth in their expression. We introduce GenTree as an efficient algorithm to learn robust rules. Experiments using GenTree with two real-world datasets show that learned rules convey more comprehensive information than previously possible.

Association Rules

"Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories." … Han

Given sets of items in an information repository, an association rule is an expression of the form Body ⇒ Head, where Body and Head are sets of items. An association rule can be stated as "Body tends to Head."

For a rule Body ⇒ Head [Support, Confidence]:

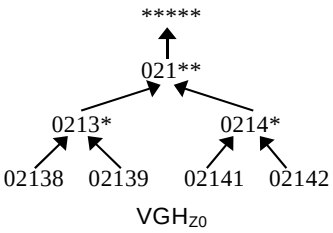
Support is the probability that both Body and Head are satisfied: Prob(Body & Head)

Confidence is the conditional probability that given Body is satisfied, Head is also satisfied: Prob(Head | Body)

Hierarchies & Robust Rules

Hierarchies provide levels of aggregation, all data types can have meaningful hierarchies. A hierarchy is based on the semantics of the attribute. It is semantically pre-defined and not automatically computed from values.

Hierarchy example: The encoding of ZIP codes 02138, 02139, etc.



Traditional association rule mining does not consider hierarchy

- Rules must have values from the same level of aggregation
- "People living in ZIP 02139, tend to be Democrats. " [Support: 19.0%, Confidence: 57.9%]

Basic generalized association rule learning is limited

- One dimensional –hierarchy on values of one attribute
- Limited expressivity, unable to learn rules "in depth"
- Redundancy in rules that are learned

Our Robust rules are better

- Values from different levels of aggregation across multiple attributes
- For the same set of tuples, maximum expression of information through most general form of Body (input) and most specific form of Head (output)

Example: "Women living in Cambridge (021**) and registered in 1970's (197*/**/**), tend to be Democrats." [Support: 2.6%, Confidence: 83.0%]

Carnegie Mellon
DATA PRIVACY LAB

Adding Semantics and Rigor to Association Rule Learning: the GenTree Approach

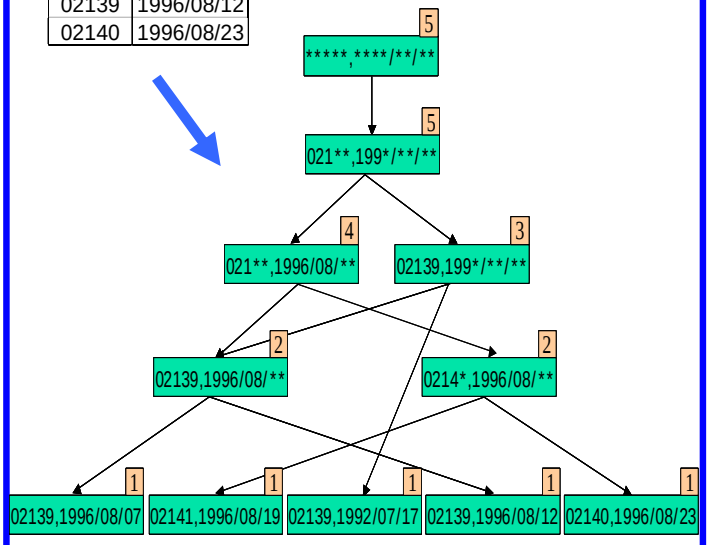
Yiheng Li and Latanya Sweeney
School of Computer Science, Carnegie Mellon University
http://privacy.cs.cmu.edu/dataprivacy/projects/gentree/

GenTree & Rule Mining

A *GenTree* is a Directed Acyclic Graph
- represents multi-dimensional cross-level generalization relations
- accounts for all data tuples in a dataset over a set of hierarchical attributes
- satisfies the properties of *completeness* and *conciseness*

Two type of nodes in GenTree
- *Leaf* node: represents a corresponding data tuple
- *Non-leaf* node:
  - represents a multi-dimensional cross-level generalization form
  - also represents the set of data tuples that can be expressed by that form
  - *root* is a special *non-leaf* node

ZIP	RegDate
02139	1996/08/07
02141	1996/08/19
02139	1992/07/17
02139	1996/08/12
02140	1996/08/23



GenTree Example: upper-right corner of each node shows the number of tuples represented.

- A node ***b*** and an ancestor ***a*** make a basic rule: *Form*(***a***) ⇒ *Form*(***b***)
- Support = |*Tuples*(*b*)| / |D|
- Confidence = |*Tuples*(*b*)| / |*Tuples*(*a*)|
- Further generalize *Form*(***a***) to most general form(s) representing the same tuple set
- Search GenTree top-down to find all robust rules satisfying specified requirements

Examples & Experimental Results

PRIOR WORK

Association Rule: "people who buy diapers tend to buy beer."
Generalized Rule: "people who buy baby products tend to buy controlled substances."

OUR NEW WORK

Robust Rule: "people who buy baby products tend to buy controlled substances, use credit cards and make purchases on Saturdays."

Dataset: Cambridge Voter List, 1997: Sampled 10,000 out of 54,805 records

- 4140 rules learned, see examples below:

	ZIP	Party	Sex	Birthdate	Regdate	Status	Tends	Zipcode	Party	Sex	Birthdate	Regdate	Status	Support	Confidence	KR
1	********	*	*	****/**/**	****/**/**	*	==>	021******	*	*	19**/**/**	19**/**/**	*	99.86%	99.86%	3
2	********	*	F	19**/**/**	19**/**/*	A	==>	021******	D	F	19**/**/**	19**/**/**	A	29.28%	65.42%	2
3	********	*	M	19**/**/**	19**/**/*	A	==>	021******	D	M	19**/**/**	19**/**/**	A	20.53%	56.62%	2
4	********	*	*	****/**/**	****/**/**	*	==>	021******	D	*	19**/**/**	19**/**/**	A	52.47%	52.47%	5
5	********	*	F	19**/**/**	197*/**/**	*	==>	0213*****	*	F	19**/**/**	197*/**/**	A	2.19%	65.37%	3
6	********	*	F	197*/**/**	1996/**/*	*	==>	021******	*	F	197*/**/**	1996/**/**	A	3.27%	99.70%	2

1	"All voters (tend to) live in Cambridge (021**), were born in the 1900's (19**/**/**) and registered to vote in the 1900's (19**/**/**)." *Commentary: The voter list for Cambridge, Massachusetts contains people who live in Cambridge and were registered to vote in the 1900's. A few people were born in the 1800's, but as shown by the rule above, almost all voters were born in the 1900's.*
2	"Female voters born in the 1900's, registered in the 1900's, and who are active voters tend to live in Cambridge and be Democrats."
3	"Male voters born in the 1900's, registered in the 1900's, and who are active voters tend to live in Cambridge and be Democrats."
4	"Voters tend to live in Cambridge, be Democrats, been born in the 1900's, registered in the 1900's, and are active voters." *Commentary: About half the voters in the Cambridge voter list are registered Democrats; the other half of the voters have no party affiliation or are registered as Republicans. This is confirmed by the support for rules 2 and 3 which together total 49.81%. In comparison, rule 4 alone, which has one of the largest knowledge rating (KR) values awarded to the dataset, states this concept succinctly.*
5	"Female voters born in the 1900's and registered in the 1970's tend to live near MIT and Harvard (0213**) and be active voters."
6	"Female voters born in the 1970's and registered in 1996 tend to live in Cambridge and be active voters."

Dataset: Voter List for ZIP 15213, totally 4,316 records

[Source: Pittsburgh Voter List 2001]

- Attributes:{*sex*, *birthdate*, *regis_date*, *party_code*, *ethnic_code*, *income*, *home_owner*, *havechild*}
- 8160 rules mined with *minsup*=2% and *minconf*=50%
- Rule Examples:
  - {*,196*/**/**,198*/**,D,*,*,*,*} ⇒ {F,196*/**/**,198*/**,D,*,*,*,*}
    "Democrats born in 1960's and registered in 1980's tend to be Female".
    [Support: 2.2%, Confidence: 52.2%, KR: 1]
  - {*, ****/**/**, 19**/**, R, W, *, D, *} ⇒ {F, 19**/**/**, 19**/**, R, W, *, D, F}
    "Republican Whites owning home tend to be females with no children."
    [Support: 2.1%, Confidence: 55.4%, KR: 3]

	Pittsburgh ($\mathcal{D}_1$)	Cambridge ($\mathcal{D}_2$)
Number of tuples	4316	10,000
Number of attributes	8	6
Number of hierarchies having height > 2	2	3
Number of robust rules learned	8160	4140
Number of rules learning in depth	167	1117

Summary results of applying GenTree to two voter lists