Exploratory Hypothesis Testing and Analysis

PI: Wong Limsoon, National University of Singapore

- Objectives:
  - Formulate “hypothesis mining” as contextualized comparative pattern mining. Develop algorithms for hypothesis mining and analysis. Build an easy-to-use system based on these algorithms.
- Novelty:
  - Hypothesis as a contextualized comparative pattern
  - P-value for controlling false-positive hypothesis
  - Comparison between hypotheses to identify actionable hypotheses, redundant hypotheses, Simpson’s paradoxes
- Scope & Deliverables:
  - Core algorithms for hypothesis generation and novel analyses described above, as well as OLAP operations for exploring hypotheses.
  - Graphical user interface (GUI) which supports basic functions for summarizing and visualizing data, visualization of discovered hypotheses, and visualization of OLAP operations

- Team Members:

Wong Limsoon (PI)

Feng Mengling (Collaborator)

Liu Guimei (RF)

Wilson Goh (RA)

Zhang Haojun (RA)

Andre Suchitra (RA)

Wang Yue (PhD Student)

Qian Jiangwen (FYP student)
Du Linxiang (FYP student)
Ruchi Bajoria (FYP student)

- Achievement #1
  - Novel formulation of hypothesis patterns as contextualized comparative patterns. E.g., $\langle$\{Race=Chinese\}, Drug=A|B, Response=positive$\rangle$
  - Novel data-driven paradigm of hypothesis generation and testing
  - Novel and efficient algorithms for generating significant hypotheses, isolating reasons behind significant hypotheses, and detecting confounding factors that form Simpson;s paradoxes with significant hypotheses
- Associated Technology
  - Implemented these algo’s into the EHTA system, the mining engine of iDIG in I^{2}R
- References
  - Liu, et al. **Towards exploratory hypothesis testing and analysis.** *ICDE 2011,* pp. 745-756
  - Liu, et al. **Supporting exploratory hypothesis testing and analysis**. *ACM KDD*, in press

Timing Performance

RUNNING TIME (MEASURED IN SECONDS) AND NUMBER OF SIGNIFICANT HYPOTHESES GENERATED. *max_pvalue is set to 0.05 on all datasets. “GenH”: time for testing all hypotheses; “AnalyzeH”: time for analysis of all significant hypotheses; “AvgAnalyzeT”: average time for analyzing a single hypothesis; “#test”: total number of tests performed; “#SignH”: #significant hypotheses with p-value≤ max_pvalue; “#SignH_BC”: #significant hypotheses with p-value≤ max_pvalue/#test (Bonferroni correction); “#SignH_FDR0.05”: #significant hypotheses when FDR is set at 0.05 (Benjamini and Hochberg’s method).*

Datasets	*min_sup*	*min_diff*	GenH	AnalyzeH	AvgAnalyzeT	#tests	#SignH	#SignH_BC	#SignH_FDR0.05
adult	500	0.05	0.42 sec	6.30 sec	0.0015 sec	5593	4258	3929	4257
adult	100	0.05	2.69 sec	37.39 sec	0.0014 sec	41738	26095	16345	25506
mushroom	500	0.1	0.67 sec	19.00 sec	0.0020 sec	16400	9323	9244	9323
mushroom	200	0.1	5.45 sec	123.47 sec	0.0020 sec	103025	61429	57798	61429
DrugTestI	20	0.5	0.06 sec	0.06 sec	0.0031 sec	3627	20	1	1
DrugTestII	20	0.5	0.08 sec	0.30 sec	0.0031 sec	4441	97	53	97

Case Study: Causal Factor Analysis

AN EXAMPLE OF A SIGNIFICANT HYPOTHESIS IDENTIFIED FROM DATASET *DrugTestII* AND POSSIBLE REASONS BEHIND IT. *SNP_OATPB_14 is a SNP at locus 14 of gene OATPB.*

ID	Context	Comparing Groups	sup	*mean*	p-value
H_1	{}	Ethnic_Group=Japanese	25	**5.256**	≤ 0.001
		Ethnic_Group=Caucasian	22	**4.750**	

(a) An example of significant hypotheses in dataset *DrugTest II*

Context	Extra Attribute	Comparing Groups	sup	*mean*
{}	SNP_OATPB_14=0	Ethnic_Group=Japanese	7	**4.749**
		Ethnic_Group=Caucasian	21	**4.785**
	SNP_OATPB_14=1	Ethnic_Group=Japanese	17	**5.489**
		Ethnic_Group=Caucasian	1	**4.009**

(b) The attribute with the highest *contribution* with respect to H_1.

Case Study: Simpson’s Paradox

(A) EXAMPLES OF SIGNIFICANT HYPOTHESES IDENTIFIED FROM DATASET *adult*. (B) A SIMPSON’S PARADOX BEHIND HYPOTHESIS H_1. *Column $P_{>50K}$ is the proportion of instances with annual income >50K.*

ID	Context	Comparing Groups	sup	$P_{>50K}$	p-value
H_1	Race = White	Occupation = Craft-repair	3694	**22.84%**	≤1.00E-08
		Occupation = Adm-clerical	3084	**14.23%**	
H_2	Race = White	Sex=Male	19174	**31.76%**	≤ 1.00E-08
		Sex=Female	8642	**11.90%**	

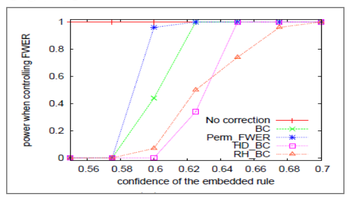
(a) Examples of significant hypotheses in dataset *adult*

Context	Extra Attribute	Comparing Groups	sup	$P_{>50K}$
Race = White	Sex = Male	Occupation = Craft-repair	3524	**23.5%**
		Occupation = Adm-clerical	1038	**24.2%**
	Sex = Female	Occupation = Craft-repair	107	**8.8%**
		Occupation = Adm-clerical	2046	**9.2%**

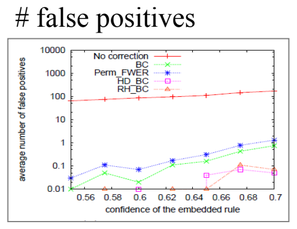
(b) A Simpson’s Paradox behind H_1.

- Achievement #2
  - Proving necessity of controlling false positives in class-association rule mining: Many spurious rules are produced if no correction is made
  - Proving that permutation-based approach is most effective in controlling false positives, and develop techniques to make it efficient
- Associated Technology
  - Implemented these techniques into ARminer

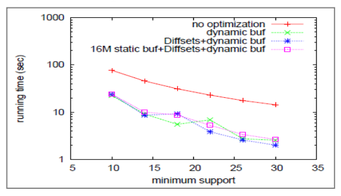
Power



false positives



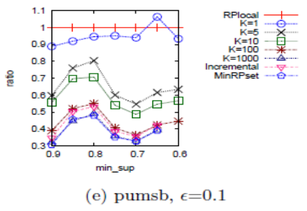
Speeding up permutation test
(i) Mine rules only once. (ii) Diffsets. (iii) Buffer p-values



- Reference
  - Liu, et al**. Controlling false positives in association rule mining**. *Proc. VLDB Endowment, 5(2):145-156, 2011*

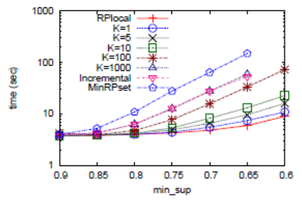
- Achievement #3
  - Finding minimum representative rule set that can 1/ represent all patterns with a minimum # of representative patterns and 2/ restore the support of all patterns with error guarantee
  - Algorithms for doing the above efficiently: MinRPset (efficiently produces the smallest solution), FlexRPset (trades solution size for higher speed)
- Associated Technology
  - Implemented these into the EHTA system, the mining engine of iDIG in I^{2}R

# of Rep Patterns



(e) pumsb, ϵ=0.1

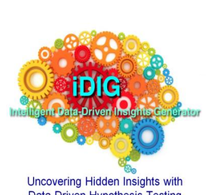
Running Time



(e) pumsb, ϵ=0.1

- Reference
  - Liu et al. **Finding minimum representative pattern sets**. *KDD 2012*, pp 51-59
  - Liu et al. **A flexible approach to finding representative pattern sets**. *IEEE TKDE*, 26(7):1562-1574, 2014

- Achievement #4
  - Association-rule visualization for exploratory data analysis
    - ❑ Relationship among rules reveal deep info of the data
    - ❑ Summarize this, with visualization, to help users understand the data and to suggest hypotheses to test
- Associated Technology
  - Implemented in AssocExplorer, the visualization engine of iDIG in I^{2}R



Uncovering Hidden Insights with Data-Driven Hypothesis Testing

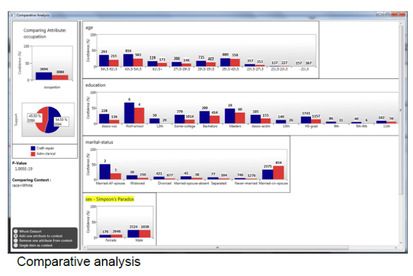
Examples

ID	Gender	Education	Occupation	Income
1	F	Bachelor	Adm-clerical	>50K
2	M	High-School	Sales	≤50K
...	...	...	...	...

An example dataset

Typical questions:

1. Which groups of people are more likely to have a high income?
2. Which attributes are important to income?
3. What is the effect of "Education" on income with respect to other attributes?
4. Women earn less than men in general. How can women have a high income?



Comparative analysis

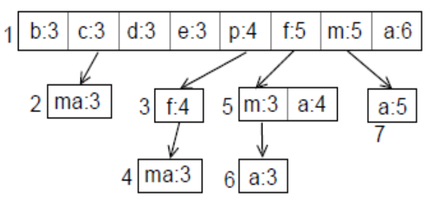
- Reference
  - Liu et al. **AssocExplorer: An association rule visualization system for exploratory data analysis**. *KDD 2012*, pp. 1536-1539

- Achievement #5
  - Management of large collections of frequent itemsets for analysis and user exploration.
  - Refinement of CPFtree to index to provide efficient exact match, subset/superset search, etc. of frequent itemsets

Example data

ID	Itemsets	ID	itemsets	ID	itemsets
1	a:6	9	ac:3	17	acm:3
2	b:3	10	af:4	18	afm:3
3	c:3	11	am:5	19	afp:3
4	d:3	12	ap:3	20	amp:3
5	e:3	13	cm:3	21	fmp:3
6	f:5	14	fm:3	22	afmp:3
7	m:5	15	fp:4		
8	p:4	16	mp:3		

CPFtree



- Reference
  - Liu et al. **A performance study of three disk-based structures for indexing and querying frequent itemsets.**. *Proc. VLDB Endowment, 6(7):505-516, 2013*

Contact- Professor Wong Limsoon (wongls@comp.nus.edu.sg)

Project URL- http://www.comp.nus.edu.sg/~wongls/projects/hypothesis/index.html

Funding- A*STAR SERC PSF grant # 102 101 0030

NUS National University of Singapore
School of Computing