Exploratory Hypothesis Testing & Analysis

Limsoon Wong

27 February 2013

NUS
National University of Singapore

Project Outline

- **Objectives**
  - Help users understand their data
  - Find actionable knowledge
- **Scope**
  - Hypothesis mining algo
  - GUI for visualization and summarization
  - Real-life applications
- **Novelty**
  - Focus on hypothesis
    - **i.e., a comparison of two samples**
  - More informative than patterns and rules
    - **Users not only get to know what is happening but also when or why it is happening**

Project Achievement #1

- **Algo's for mining, testing, & analyzing hypothesis**
  - Novel formulation of a hypothesis into context, comparing attribute, and target attribute
    - **E.g., ⟨{Race=Chinese}, Drug=A|B, Response=positive⟩**
  - Novel algo for exploratory hypothesis testing
  - Novel algo for hypothesis analysis
- **Implemented these algo's into the EHTA system, the mining engine of iDIG in I2R, which can help users**
  - Identify significant hypotheses
  - Isolate reasons behind significant hypotheses
  - Find confounding factors that form Simpson's Paradoxes with discovered significant hypotheses

Liu, et al. Towards exploratory hypothesis testing and analysis. *Proc ICDE 2011*, pages 745-756

Outline

- Background
- Problem definition
- Algorithms
- Experiments
- Related work
- Summary and discussion

Background

- **A hypothesis compares two or more groups**
  - Do smokers have higher cancer rates than non-smokers?
  - Are children more vulnerable to H1N1 flu than adults?
- **Statistical hypothesis testing**
  - Test whether a hypothesis is supported by data using statistical methods

Conventional Hypothesis Generation

- **Postulate a hypothesis**
  - Is drug A more effective than drug B?
- **How?**
  - Collect data and eye ball a pattern!

PID	Race	Sex	Age	Smoke	Stage	Drug	Response
1	Caucasian	M	45	Yes	1	A	positive
2	Chinese	M	40	No	2	A	positive
3	African	F	50	Yes	2	B	negative
…	…	…	…	…	…	…	…
N	Caucasian	M	60	No	2	B	negative

P-Value

- **Use statistical methods to decide whether a hypothesis “Is drug A more effective than drug B? ” is supported by data**
  - E.g., χ2-test

	Response= positive	Response= Negative	Proportion of positive responses
Drug=A	890	110	89%
Drug=B	830	170	83%

- **p-value = 0.0001**
  - Prob of observed diff betw the two drugs given assumption that the they have same effect

Limitations of Conventional Approach

- **Hypothesis-driven**
  - Scientist has to think of a hypothesis first
  - Allow just a few hypotheses to be tested at a time

- **So much data have been collected …**
  - No clue on what to look for
  - Know something; but do not know all
  - Impossible to inspect so much data manually

$\Rightarrow$ **Exploratory hypothesis testing in a data-driven manner**

Exploratory Hypothesis Testing

- **Data-driven hypothesis testing**
  - Have a dataset but dunno what hypotheses to test
  - Use computational methods to automatically formulate and test hypotheses from data

- **Problems to be solved:**
  - How to formulate hypotheses?
  - How to automatically generate & test hypotheses?

Outline

- **Background**
- **Problem definition**
- **Algorithms**
- **Experiments**
- **Related work**
- **Summary and discussion**

Formulation of a Hypothesis

- **"For Chinese, is drug A better than drug B?"**

- **Three components of a hypothesis:**
  - Context (under which the hypothesis is tested)
    - **Race: Chinese**
  - Comparing attribute
    - **Drug: A or B**
  - Target attribute/target value
    - **Response: positive**

- **⟨{Race=Chinese}, Drug=A|B, Response=positive⟩**

Testing a Hypothesis

- **⟨{Race=Chinese}, Drug=A|B, Response=positive⟩**

- **To test this hypothesis we need info:**
  - N^A =support({Race=Chinese, Drug=A})
  - N^A_{pos} =support({Race=Chinese, Drug=A, Res=positive})
  - N^B =support({Race=Chinese, Drug=B})
  - N^B_{pos} =support({Race=Chinese, Drug=B , Res=positive})

context	Comparing attribute	response= positive	response= negative
{Race=Chinese}	Drug=A	N^A_{pos}	$N^A - N^A_{pos}$
	Drug=B	N^B_{pos}	$N^B - N^B_{pos}$

$\Rightarrow$ **Frequent pattern mining**

Significance of Observed Diff

- **When a single hypothesis is tested, a p-value of 0.05 is recognized as low enough**
  - If we test 1000 hypotheses, ~50 hypotheses will pass the 0.05 threshold by random chance!

- **Control false positives**
  - Bonferroni's correction
    - **Family-Wise Error Rate: Prob of making one or more false discoveries**
  - Benjamini and Hochberg's method
    - **False Discovery Rate: Proportion of false discoveries**
  - Permutation method

Need for Hypothesis Analysis

- **Exploration is not guided by domain knowledge**

  ⇒Spurious hypotheses has to be eliminated

- **Reasons behind significant hypotheses**
  - Find attribute-value pairs that change the diff a lot
    - **DiffLift: How much diff betw the two groups is lifted**
    - **Contribution: Freq of attribute-value pairs**

DEFINITION 3 (*DiffLift*$(A{=}v|H)$). *Let* $H = \langle P, A_{diff} = \{v_1, v_2\}, A_{target}, v_{target}\rangle$ *be a hypothesis,* A_{target} *be categorical,* $P_1 = P \cup \{A_{diff} = v_1\}$ *and* $P_2 = P \cup \{A_{diff} = v_2\}$ *be the two sub-populations of* H, $A = v$ *be an item not in* H*, that is,* $A \neq A_{diff}$, $A \neq A_{target}$ *and* $A = v \notin P$*. After adding item* $A = v$ *to* H*, we get two new sub-populations:* $P'_1 = P_1 \cup \{A = v\}$ *and* $P'_2 = P_2 \cup \{A = v\}$*. The lift of difference after adding* $A = v$ *to* H *is defined as* $DiffLift(A{=}v|H) = \frac{p'_1 - p'_2}{p_1 - p_2}$*, where* p_i *is the proportion of* v_{target} *in sub-population* P_i*, and* p'_i *is the proportion of* v_{target} *in sub-population* P'_i*,* $i{=}1, 2$.

DEFINITION 6 (*Contribution*$(A = v|H)$). *Let* H *be a hypothesis,* $A = v$ *be an attribute value not in* H*,* P_1 *and* P_2 *be the two sub-populations of* H*,* P'_1 *and* P'_2 *be the two sub-populations after adding* $A = v$ *to* H*. The contribution of* $A = v$ *to* H *is defined as* $Contribution(A = v|H) = \frac{\frac{n'_1}{n_1}(p'_1 - p_1) - \frac{n'_2}{n_2}(p'_2 - p_2)}{p_1 - p_2}$*, where* p_i *is the proportion of* v_{target} *in sub-population* P_i*, and* p'_i *is the proportion of* v_{target} *in sub-population* P'_i*,* $i = 1, 2$.

Spurious Hypotheses

	response= positive	response= negative	proportion of positive response
Drug=A	890	110	89.0%
Drug=B	830	170	83.0%
Drug=A, Stage=1	800	80	90.9%
Drug=B, Stage=1	190	10	95%
Drug=A, Stage=2	90	30	75%
Drug=B, Stage=2	640	160	80%

- **Simpson's Paradox**
  - "Stage" has assoc w/ both "drug" & "response":
    - **Doc's tend to give drug A to patients at stage 1, & drug B to patients at stage 2**
    - **Patients at stage 1 are easier to cure than patients at stage 2**
  - Attribute "stage" is called a confounding factor

Reasons Behind Significant Hypotheses

	Failure rates
Product A	4%
Product B	2%
Product A, time-of-failure=loading	6.0%
Product B, time-of-failure=loading	1.9%
Product A, time-of-failure=in-operation	2.1%
Product B, time-of-failure=in-operation	2.1%
Product A, time-of-failure=output	2.0%
Product B, time-of-failure=output	1.9%

- **Problem is narrowed down**
  - Product A has exceptionally higher drop rate than product B only at the loading phase

Problem Statement: Exploratory Hypothesis Testing

- **Given**
  - Dataset D, min_sup, max_pvalue, min_diff
  - $A_{target} = v_{target}$
  - $\mathcal{A}_{grouping}$: context/comparing attributes

- **Find all $H = \langle P, A_{diff} = v_1|v_2, A_{target} = v_{target} \rangle$**
  - $A_{diff} \in \mathcal{A}_{grouping}$ & $\forall(A=v)$ in P, $A \in \mathcal{A}_{grouping}$
  - $sup(P_i) \geq$ min_sup, where $P_i = P \cup \{A_{diff} = v_i\}$, i=1, 2
  - p-value(H) $\leq$ max_pvalue
  - $|p_1 - p_2| \geq$ min_diff, where p_i is proportion of v_{target} in sub-population P_i, i=1, 2

Problem Statement: Hypothesis Analysis

- **Given a significant hypothesis H, generate the following info for further analysis**
  - Simpson's Paradoxes formed by H with attributes not in H
  - List of attribute-value pairs not in H ranked in descending order of DiffLift(A=v|H) and Contribution(A=v|H)
  - List of attributes not in H ranked in descending order of DiffLift(A|H) and Contribution(A|H)

Outline

- **Background**
- **Problem definition**
- **Algorithms**
- **Experiments**
- **Related work**
- **Summary and discussion**

Algo for Exploratory Hypothesis Testing

- **A hypothesis is a comparison betw two or more sub-populations, and each sub-populations is defined by a pattern**

- **Step 1: Use freq pattern mining to enumerate large sub-populations and collect their statistics**
  - Stored in the CFP-tree structure, which supports efficient subset/superset/exact search
- **Step 2: Pair sub-populations up to form hypotheses, and then calculate their p-values**
  - Use each freq pattern as a context
  - Search for immediate supersets of the context patterns, and then pair these supersets up to form hypotheses

Algo for Hypothesis Analysis

- **Given a hypothesis H**
  - To check whether H forms a Simpson's Paradox with an attribute A,
    - **add values of A to context of H**
    - **re-calculate the diff betw the two sub-populations**
  - To calculate DiffLift and Contribution of an attribute-value pair A=v,
    - **add A=v to context of H**
    - **re-calculate the diff**

- **All can be done via immediate superset search**

Outline

- **Background**
- **Problem definition**
- **Algorithms**
- **Experiments**
- **Related work**
- **Summary and discussion**

Experiment Settings

- **PC configurations**
  - 2.33Ghz CPU, 3.25GB memory, Windows XP
- **Datasets:**
  - mushroom, adult: UCI repository
  - DrugTestI, DrugTestII: study assoc betw SNPs in several genes & drug responses.

Datasets	#instances	#continuous attributes	#categorical attributes	A_{target}/v_{target}
adult	48842	6	9	class=>50K (nominal)
mushroom	8124	0	23	class=poisonous (nominal)
DrugTestI	141	13	74	logAUCT (continuous)
DrugTestII	138	13	74	logAUCT (continuous)

Running Time

- **Three phases**
  - Frequent pattern mining
  - Hypothesis generation
  - Hypothesis analysis

Datasets	min_sup	min_diff	GenH	AnalyzeH	AvgAnalyzeT	#tests	#signH
adult	500	0.05	0.42 s	6.30 s	0.0015 s	5593	4258
adult	100	0.05	2.69 s	37.39 s	0.0014 s	41738	26095
mushroom	500	0.1	0.67 s	19.00 s	0.0020 s	16400	9323
mushroom	200	0.1	5.45 s	123.47 s	0.0020 s	103025	61429
DrugTestI	20	0.5	0.06 s	0.06 s	0.0031 s	3627	20
DrugTestII	20	0.5	0.08 s	0.30 s	0.0031 s	4441	97

max_pvalue = 0.05

Case Study: Adult Dataset

Context	Comparing Groups	sup	$P_{class=>50K}$	p-value
Race =White	Occupation = Craft-repair	3694	22.84%	1.00×10^{-19}
	Occupation = Adm-clerical	3084	14.23%	

- **Simpson's Paradox**

Context	Extra attribute	Comparing Groups	sup	$P_{class=>50K}$
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%

Summary

- **Formulated the exploratory hypothesis testing and analysis problem**
  - Complementary to conventional hypothesis testing
  - Overcome human oversights & limitations
  - Further analysis:
    - **Narrow down the problem**
    - **Find Simpson's Paradox**
- **Proposed a data mining approach for this**
  - Efficient

What's next?

- **Controlling false positive rate**
  - Bonferroni's correction
  - Benjamini and Hochberg's method
  - Permutation test

- **Concise representations of hypotheses**
  - freq patterns & hypotheses have lots of redundancy

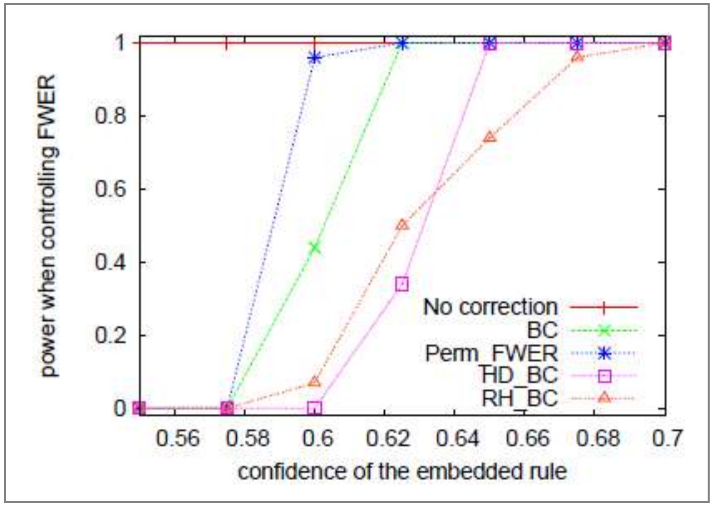
- **Organization & presentation of hypotheses**
  - Visualization
  - Summarization

Project Achievement #2

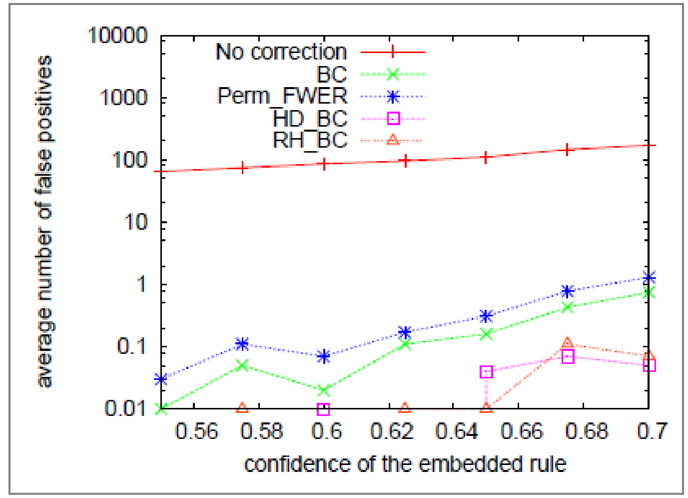
- **Control false positives in class-association rule mining**
  - Large # of rules being tested. Rules not representing real effect can satisfy the constraints purely by random chance
- **Three approaches to control false positives**
  - Direct adjustment, e.g., Bonferroni's
  - Permutation-based p-value
  - Holdout approach
- **We show that**
  - Many spurious rules are produced if no correction is made
  - These approaches can control false positives effectively
  - Permutation-based approach is most effective, but costly
  - Techniques to make permutation-based approach efficient

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

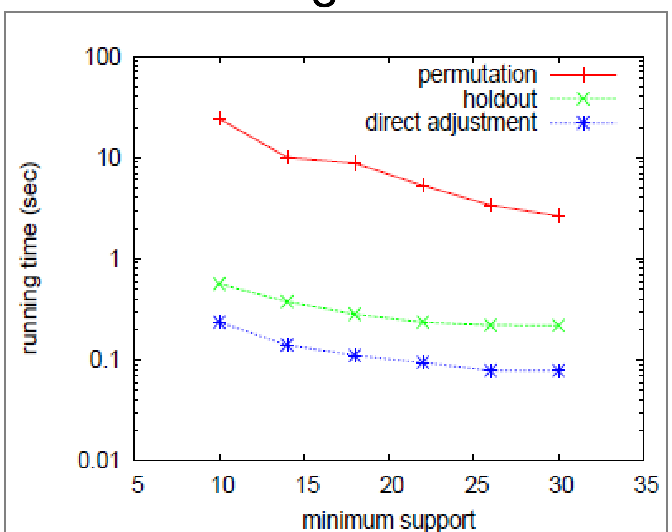
Power



false positives

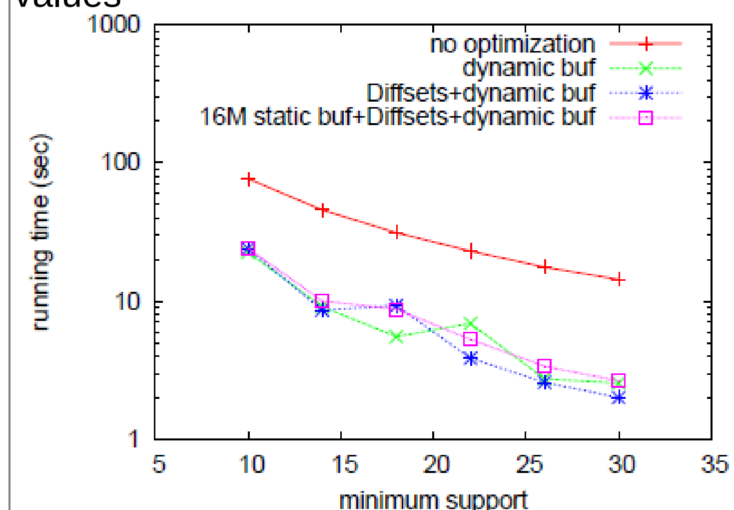


Running Time



Speeding up permutation test

(i) Mine rules only once. (ii) Diffsets. (iii) Buffer p-values



Project Achievement #3

- **Finding minimum representative rule sets**
  - Freq pattern mining often produces many freq patterns
  - Difficult to understand the generated patterns

- **Challenges**
  - Produce a minimum # of representative patterns
  - Can restore the support of all patterns with error guarantee
  - Do the above efficiently

- **We develop MinRPset and FlexRPset**
  - MinRPset always efficiently produces the smallest solution
  - FlexRPset can trade solution size for even higher speed

Liu, et al. **Finding minimum representative rule sets**. *Proc KDD2012*, pages 51-59

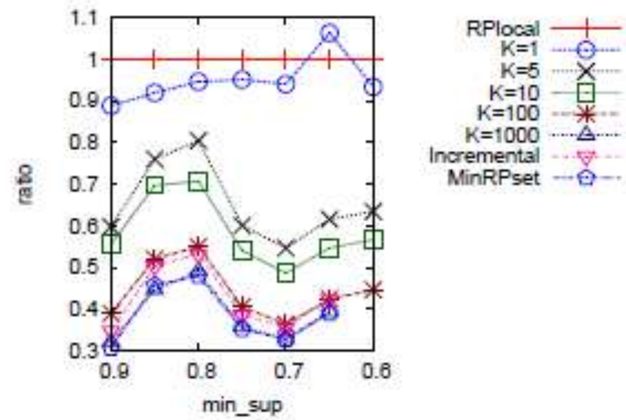
Definition 1 ($D(X_1, X_2)$). *Given two patterns X_1 and X_2, the distance between them is defined as* $D(X_1, X_2) = 1 - \frac{|T(X_1) \cap T(X_2)|}{|T(X_1) \cup T(X_2)|}$.

Definition 2 (ϵ-covered). *Given a real number $\epsilon \in [0, 1]$ and two patterns X_1 and X_2, we say X_1 is ϵ-covered by X_2 if $X_1 \subseteq X_2$ and $D(X_1, X_2) \leq \epsilon$.*

In the above definition, condition $X_1 \subseteq X_2$ ensures that the two patterns have similar items, and condition $D(X_1, X_2) \leq \epsilon$ ensures that the two patterns have similar supporting transaction sets and similar support. Based on the definition, a pattern ϵ-covers itself.

Lemma 1. *Given two patterns X_1 and X_2, if pattern X_1 is ϵ-covered by pattern X_2 and we use $supp(X_2)$ to approximate $supp(X_1)$, then the relative error $\frac{supp(X_1) - supp(X_2)}{supp(X_1)}$ is no larger than ϵ.*

of Rep Patterns

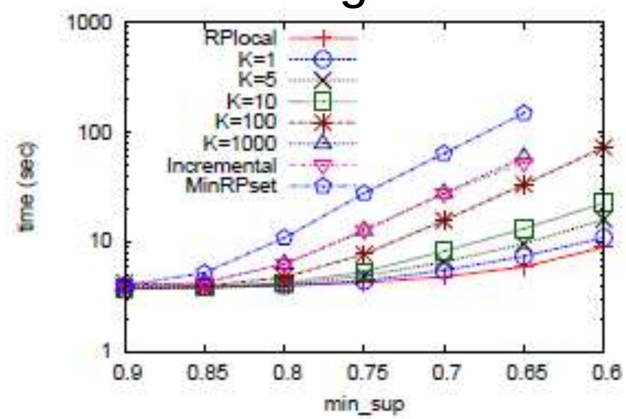


(e) pumsb, ϵ=0.1

Table 4: Running time of MinRPset with and without the early termination technique.

dataset	*min_sup*	ϵ	W/O(sec)	With(sec)	ratio
accidents	0.2	0.1	12.139	2.406	19.8%
accidents	0.2	0.05	10.280	1.640	16.0%
chess	0.3	0.1	323.964	48.107	14.8%
chess	0.3	0.05	240.312	22.392	9.3%
connect	0.2	0.1	104.444	15.014	14.4%
connect	0.2	0.05	88.492	5.625	6.4%
mushroom	0.001	0.2	3.312	0.312	9.4%
mushroom	0.001	0.1	0.281	3.266	8.6%
mushroom	0.001	0.05	0.265	3.266	8.1%
pumsb	0.6	0.1	160.670	242.33	66.3%
pumsb	0.6	0.05	34.687	106.388	32.6%
pumsb_star	0.1	0.1	109.796	24.904	22.7%
pumsb_star	0.1	0.05	88.148	13.934	15.8%

Running Time



(e) pumsb, ϵ=0.1

Project Achievement #4

- **Association rule visualization system 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**

Main Features

1. Visual information-seeking mantra: overview first, zoom and filter, details on demand.
2. Use coloring to deliver information effectively.
3. If users find a rule interesting, they can explore related rules to have a deeper understanding of the rule.
4. Rules with similar item composition but very different statistics may represent inexpensive actions that we can take to make a big change. Our system allow users to inspect such rules under various contexts to isolate the key factors that contribute to the difference.

- **Techniques implemented in AssocExplorer, the visualization engine of iDIG in I^{2}R**

Liu, et al. **AssocExplorer: An association rule …** *Proc KDD2012*, pages 1536-1539

Uncovering Hidden Insights with Data-Driven Hypothesis Testing

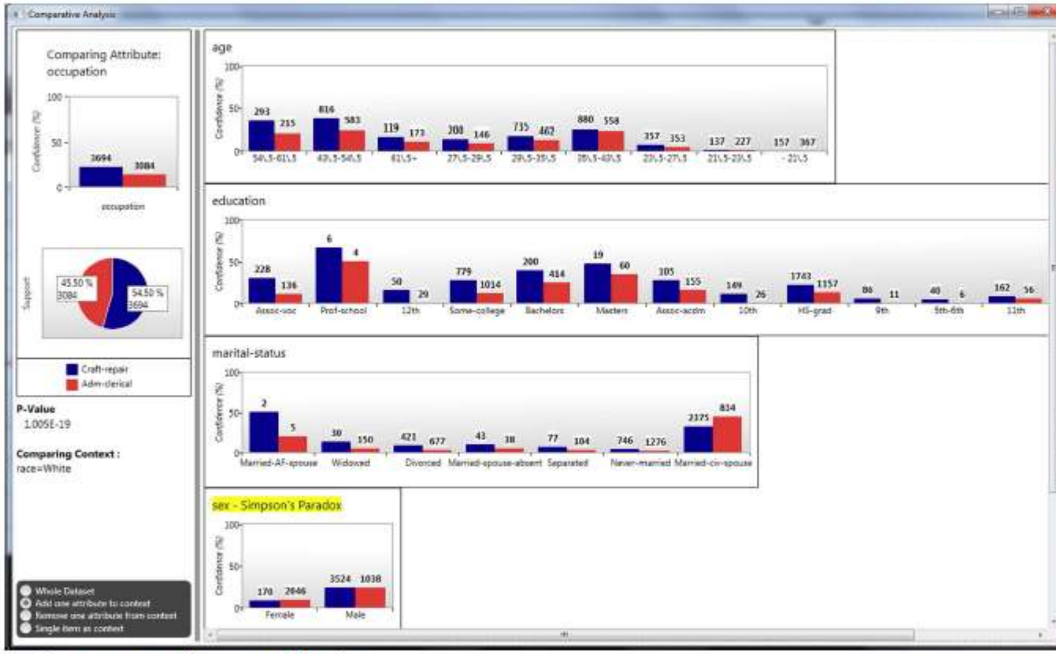
Examples

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

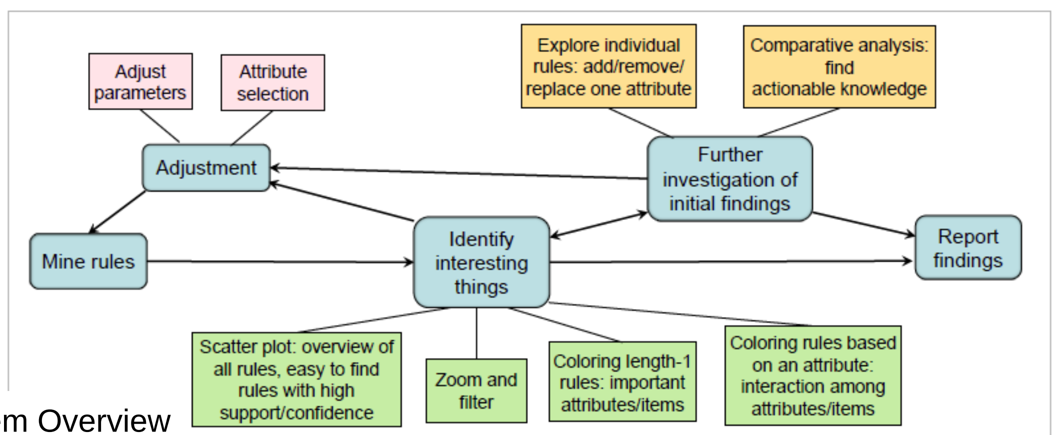
An example dataset

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



System Overview

Summary of Project Results

- **Deliverables achieved:**
  - Algorithms for
    - **Exploratory hypothesis testing and analysis (EHTA)**
    - **Selecting minimum representative rules**
    - **Efficiently controlling false positives**
    - **Visualization system for exploratory data analysis (AssocExplorer)**
  - EHTA & AssocExplorer put into iDIG at I^{2}R

- **Capabilities developed:**
  - Expertise in a novel aspect of analytics

Acknowledgements

- *Liu Guimei*
- *Zhang Hao Jun*
- *Andre Suchitra*
- *Feng Mengling*

Agency for Science, Technology and Research
SINGAPORE

Publications

- Liu, et al. **Towards Exploratory Hypothesis Testing and Analysis**. ICDE2011, pages 745-756
- Liu, et al. **Controlling False Positives in Association Rule Mining**. *Proc VLDB Endow*, 5(2):145-156, 2011
- Liu, et al. **Finding Minimum Representative Pattern Sets**. KDD2012, pages 51-59
- Liu, et al. **AssocExplorer: An association rule visualization system for exploratory data analysis**. KDD2012, pages 1536-1539
- Liu, et al. **A flexible approach to finding representative pattern sets**. *IEEE Trans Knowledge and Data Eng*, accepted.