Exploratory Hypothesis Testing & Analysis

Limsoon Wong
27 February 2013
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., $\langle\{\text{Race}=\text{Chinese}\}, \ \text{Drug}=A|B, \ \text{Response}=\text{positive}\rangle$
  – 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

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!

<table>
<thead>
<tr>
<th>PID</th>
<th>Race</th>
<th>Sex</th>
<th>Age</th>
<th>Smoke</th>
<th>Stage</th>
<th>Drug</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Caucasian</td>
<td>M</td>
<td>45</td>
<td>Yes</td>
<td>1</td>
<td>A</td>
<td>positive</td>
</tr>
<tr>
<td>2</td>
<td>Chinese</td>
<td>M</td>
<td>40</td>
<td>No</td>
<td>2</td>
<td>A</td>
<td>positive</td>
</tr>
<tr>
<td>3</td>
<td>African</td>
<td>F</td>
<td>50</td>
<td>Yes</td>
<td>2</td>
<td>B</td>
<td>negative</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>N</td>
<td>Caucasian</td>
<td>M</td>
<td>60</td>
<td>No</td>
<td>2</td>
<td>B</td>
<td>negative</td>
</tr>
</tbody>
</table>
P-Value

• Use statistical methods to decide whether a hypothesis “Is drug A more effective than drug B? ” is supported by data
  – E.g., \( \chi^2 \)-test

<table>
<thead>
<tr>
<th></th>
<th>Response= positive</th>
<th>Response= Negative</th>
<th>Proportion of positive responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug=A</td>
<td>890</td>
<td>110</td>
<td>89%</td>
</tr>
<tr>
<td>Drug=B</td>
<td>830</td>
<td>170</td>
<td>83%</td>
</tr>
</tbody>
</table>

• p-value = 0.0001
  – Prob of observed diff betw the two drugs given assumption that 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

⇒ **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

• \{\text{Race}=\text{Chinese}, \text{Drug}=A|B, \text{Response}=\text{positive}\}
Testing a Hypothesis

• \(\{\text{Race}=\text{Chinese}\}, \ \text{Drug}=A|B, \ \text{Response}=\text{positive}\) 

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

\[\Rightarrow \text{Frequent pattern mining}\]
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**
  \[ \Rightarrow \text{Spurious hypotheses has to be eliminated} \]

- **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
Spurious Hypotheses

- **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

<table>
<thead>
<tr>
<th>Drug, Stage</th>
<th>response=positive</th>
<th>response=negative</th>
<th>proportion of positive response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug=A</td>
<td>890</td>
<td>110</td>
<td>89.0%</td>
</tr>
<tr>
<td>Drug=B</td>
<td>830</td>
<td>170</td>
<td>83.0%</td>
</tr>
<tr>
<td>Drug=A, Stage=1</td>
<td>800</td>
<td>80</td>
<td>90.9%</td>
</tr>
<tr>
<td>Drug=B, Stage=1</td>
<td>190</td>
<td>10</td>
<td>95%</td>
</tr>
<tr>
<td>Drug=A, Stage=2</td>
<td>90</td>
<td>30</td>
<td>75%</td>
</tr>
<tr>
<td>Drug=B, Stage=2</td>
<td>640</td>
<td>160</td>
<td>80%</td>
</tr>
</tbody>
</table>
## Reasons Behind Significant Hypotheses

<table>
<thead>
<tr>
<th></th>
<th>Failure rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product A</td>
<td>4%</td>
</tr>
<tr>
<td>Product B</td>
<td>2%</td>
</tr>
<tr>
<td>Product A, time-of-failure=loading</td>
<td>6.0%</td>
</tr>
<tr>
<td>Product B, time-of-failure=loading</td>
<td>1.9%</td>
</tr>
<tr>
<td>Product A, time-of-failure=in-operation</td>
<td>2.1%</td>
</tr>
<tr>
<td>Product B, time-of-failure=in-operation</td>
<td>2.1%</td>
</tr>
<tr>
<td>Product A, time-of-failure=output</td>
<td>2.0%</td>
</tr>
<tr>
<td>Product B, time-of-failure=output</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

- **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
  - $A_{\text{target}} = v_{\text{target}}$
  - $A_{\text{grouping}}$: context/comparing attributes

• **Find all** $H = \langle P, A_{\text{diff}} = v_1 | v_2, A_{\text{target}} = v_{\text{target}} \rangle$
  - $A_{\text{diff}} \in A_{\text{grouping}}$ & $\forall (A = v)$ in $P$, $A \in A_{\text{grouping}}$
  - $sup(P_i) \geq \text{min}_\text{sup}$, where $P_i = P \cup \{A_{\text{diff}} = v_i\}$, $i = 1, 2$
  - $p$-value$(H) \leq \text{max}_\text{pvalue}$
  - $|p_1 - p_2| \geq \text{min}_\text{diff}$, where $p_i$ is proportion of $v_{\text{target}}$
    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 between two or more sub-populations, and each sub-population 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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#instances</th>
<th>#continuous attributes</th>
<th>#categorical attributes</th>
<th>$A_{\text{target}}/v_{\text{target}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult</td>
<td>48842</td>
<td>6</td>
<td>9</td>
<td>class=&gt;50K (nominal)</td>
</tr>
<tr>
<td>mushroom</td>
<td>8124</td>
<td>0</td>
<td>23</td>
<td>class=poisonous (nominal)</td>
</tr>
<tr>
<td>DrugTestI</td>
<td>141</td>
<td>13</td>
<td>74</td>
<td>logAUCT (continuous)</td>
</tr>
<tr>
<td>DrugTestII</td>
<td>138</td>
<td>13</td>
<td>74</td>
<td>logAUCT (continuous)</td>
</tr>
</tbody>
</table>
Running Time

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

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

max_pvalue = 0.05
Case Study: Adult Dataset

<table>
<thead>
<tr>
<th>Context</th>
<th>Comparing Groups</th>
<th>sup</th>
<th>$P_{\text{class=&gt;50K}}$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race = White</td>
<td>Occupation = Craft-repair</td>
<td>3694</td>
<td>22.84%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occupation = Adm-clerical</td>
<td>3084</td>
<td>14.23%</td>
<td>$1.00 \times 10^{-19}$</td>
</tr>
</tbody>
</table>

- **Simpson’s Paradox**

<table>
<thead>
<tr>
<th>Context</th>
<th>Extra attribute</th>
<th>Comparing Groups</th>
<th>sup</th>
<th>$P_{\text{class=&gt;50K}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race = White</td>
<td>Sex = Male</td>
<td>Occupation = Craft-repair</td>
<td>3524</td>
<td>23.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Occupation = Adm-clerical</td>
<td>1038</td>
<td>24.2%</td>
</tr>
<tr>
<td></td>
<td>Sex = Female</td>
<td>Occupation = Craft-repair</td>
<td>107</td>
<td>8.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Occupation = Adm-clerical</td>
<td>2046</td>
<td>9.2%</td>
</tr>
</tbody>
</table>
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

Power

Running Time

# false positives

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

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. \((\epsilon\text{-covered})\). Given a real number \(\epsilon \in [0, 1]\) and two patterns \(X_1\) and \(X_2\), we say \(X_1\) is \(\epsilon\)-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 \(\epsilon\)-covers itself.

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

---

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

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

- Techniques implemented in AssocExplorer, the visualization engine of iDIG in I²R

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Education</th>
<th>Occupation</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>Bachelor</td>
<td>Adm-clerical</td>
<td>&gt;50K</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>High-School</td>
<td>Sales</td>
<td>≤50K</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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?
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 I²R

• Capabilities developed:
  – Expertise in a novel aspect of analytics
Acknowledgements

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

Publications