Report for 05-Supporting Exploratory Hypothesis Testing and Analysis

Xinyu Chen

February 22, 2017

1 Summary

This paper presented a framework for automatic generating and testing hypotheses from a dataset. The authors proposed some metrics for comparison between hypotheses, so that users can carry out analysis and extract more interesting knowledge from the generated hypotheses. The paper first gave definitions for *Pattern* as a set of attributes derived from frequent pattern mining. *Hypotheses* consist *context*, *comparing* attributes and *target* attribute. Given these attributes from a hypothesis, they calculated χ^2 test statistics as the $p - value$s for subpopulations in question. This is how they test and analyze these hypotheses.

The framework is built upon a *Frequent Pattern Mining* techniques. They first find out some frequent patterns with the support greater than a min_sup parameter as *context* attributes. After that, they add *comparing* attributes to the *context* to form new hypotheses. As they adding more attributes, they check and prune the hypotheses to avoid similar and uninteresting hypotheses. They did experiments on the UCI machine learning datasets and got interesting results.

2 Key Takeaway

The paper proposed an interesting extension for *Frequent Pattern Mining*. They treated *pattern*s as *hypotheses* and measure the differences in statistics

before and after introducing new attributes to the patterns. Although the complexity seems to be high, this is an interesting approach. I like how they define the problem and how they design some metrics to measure the differences. I think they proved well to illustrate that their algorithm works.

3 Discussions

- *statistic test and p-value.* The authors seems focused on the χ^2 test and get the p-value from this test. I guess this is because they argued that only categorical attributes are interesting. I think there are other interesting tests should be applied in hypothesis testing.

- *categorical attributes.* The authors demonstrated their algorithm mainly on categorical attributes. They also focused on *comparing* attribute that has two values v_1, v_2. This is convenient for χ^2 test, but there are other interesting and important tests for continuous attributes.

- *influence metrics.* The *DiffLift* is an interesting metrics they designed to check the influences of a *comparing* attribute. They didn't mention but it may sometime generate divide by zero error. They chose this because of simplicity. Maybe there are some other statistics or numbers they can use to measure the influences.

- *Add representative Hypotheses.* The algorithm1 showed they pick all pairs of patterns from G_i. I think they meant to pick one pattern, like $P \cup A_i$, with $A_i = v_1|v_2$, such that they can compare the subpopulation of P_1, P_2. I guess they did not really take pairs of (G_i, G_j). This is confusing.

- *Time Complexity.* They authors claimed the time complexity to generate hypotheses for one frequent pattern is $\Sigma_{A \in P}|Dom(A)| - 1$. Assume the minimum value of $|Dom(a)| - 1 = d$, the lower bond of this Σ should be $|P|d$. Due to the number of $|P|$ and the number of frequent patterns, this is an expensive algorithm when the dimensionality increases.