

Data Mining Algorithm results.

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2005-08/msg00067.html>

- *From:* tdavidge@xxxxxxxxxxx
 - *Date:* 3 Aug 2005 19:32:14 -0700
-

People,

Please excuse my ignorance, but I need some clarification on various results specifically from Association and Sequential Pattern data mining algorithms. Please be aware I am a blithering idiot when it comes to anything other than the most basic understanding of statistical tests. That said, I understand the concepts of support and confidence as provided by each algorithm specifically. What I do not understand however are the lift and p-values provided with the results.

Each of these tools provide a rule_body [A], rule_head [C] together with support and lift. In the case of Association algorithm, confidence is further provided. Each also provides p-value.

Could some kind soul please elaborate [in english] on how I would interpret the lift and p-values and in conjunction with support / confidence apply that to the results.

If for example the rules are mined over 1 or two day periods, then I already calculate various values for these rules once mined. If support is x, then I know the frequency of the rule based on the number of transactions so I can figure days_rule_occurred and I can also calculate days_rule_head and days_rule_body.

These rules are returning p-values in the range of 19.19 – 37.76 [I do not know if this is a percentage or not] with support of say 49.1803%, confidence of 96.77% and lift of 1.64. Everything I have read points to p-values of less than .05 by comparison.

I have no idea how to calculate a null hypothesis or if one is even required, however I know that there is a 'null rule' that applies being the support for the consequent [or rule_head]. If for example you are mining over a period of 100 days and the rule head occurs 60 times in a particular fashion then my null rule for that consequent is 60%. Does this have anything at all to do with the null hypothesis ?

Once you have these rules [and there are a great many 000's of them] how does one determine the most statistically relevant or predictive ones from those that are not. Are there any further tests one can do on

Data Mining Algorithm results.

such rules to further gain insight ? Chisq ofr example, if there is then what data is required to perform such a test.

One last question, during mining we can see itemsets being generated that also have support, p-value and lift [no confidence of course]. How does one interpret those values to the final values that make up the actual rule ?

Many thanks for any assistance you may be able to offer.

P

.

-
- *Follow-Ups:*
 - ◆ *Re: Data Mining Algorithm re*