
Subject: Re: Chi-square decision trees

Posted by [James Kuyper](#) on Fri, 19 Apr 2002 17:05:50 GMT

[View Forum Message](#) <> [Reply to Message](#)

Dick Jackson wrote:

> Hi James,
>
> "James Kuyper" <kuyper@gscmail.gsfc.nasa.gov> wrote in message
> news:3CC030E0.9010302@gscmail.gsfc.nasa.gov...
>
>> There's a standard dataset characterization technique I used a couple
>> of decades ago, and I want to use it again, and I can't remember the
>> name of the technique.
>>
>> The context is that you have a discrete dependent variable, and a large
>> number of discrete independent variables. [...]
>>
>> Each basic step of the process involved choosing the particular variable
>> that had the most significant chi-squared value. Then, the process would
>> repeat in a hierarchical fashion on each subset determined by that
>> variable. [...]
>>
>> Does anyone recognise the technique I'm describing? Do you remember what
>> the name is? Is there an IDL routine that implements it?
>
>
> The ID3 (Iterative Dichotomizer - 3) method of Ross Quinlan may be what
> you're thinking of, although it's usually described in terms of 'information
> content' rather than 'chi-squared value', but the difference may be moot.
> It's also possible to use this method for continuous variables, with the
> extra trick of finding a split point.
>
> I once gave a talk on this method to a group of colleagues when I was doing
> work mainly in Lisp, and I had a pretty nice graphical implementation in
> object-oriented Macintosh Common Lisp. I don't know of any IDL code for it,
> but it shouldn't be too hard to do, though.
>
> I found this summary of the method through Google:
> <http://www.dcs.napier.ac.uk/~peter/vldb/dm/node11.html>

I'm positive that this is a different algorithm than the one I was talking about. It may be an equivalent one; that's hard to tell without careful analysis. It may be better; the chi-squared criterion sounded a bit ad-hoc to me; the information-theoretic derivation of this algorithm seems better-founded. However, as long as it does what it sounds like it does, I'd be willing to at least try it out.

However, I noticed that the web page had no links to an actual implementation. It mentioned a commercial package, but that doesn't help much. My current need for this tool has essentially no budget behind it. If the tool's not hidden away in one of the libraries we already have installed here (such as the IMSL or IDL libraries), I have to settle for a freeware solution, or write it myself (and that has to be low cost, too - I couldn't afford to put in more than a day or two on it). Aren't budgets fun! :-)
