Subject: Re: Chi-square decision trees
Posted by Dick Jackson on Fri, 19 Apr 2002 16:38:12 GMT
View Forum Message <> Reply to Message

Hi James,

"James Kuyper" <kuyper@gscmail.gsfc.nasa.gov> wrote in message
news:3CC030E0.9010302@gscmail.gsfc.nasa.gov...
> Theres'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 hierarchial 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

Cheers,
--
-Dick

Dick Jackson / dick@d-jackson.com
D-Jackson Software Consulting / http://www.d-jackson.com
Calgary, Alberta, Canada / +1-403-242-7398 / Fax: 241-7392