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

Subject: Re: Chi-square decision trees

Dick Jackson wrote:

> 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

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

implmentation. 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! :-(

Subject: Re: Chi-square decision trees
Posted by James Kuyper on Fri, 19 Apr 2002 21:22:10 GMT
View Forum Message <> Reply to Message

Dick Jackson wrote:

> "James Kuyper" <kuyper@gscmail.gsfc.nasa.gov> wrote in message
> news:3CC04E6E.7060304@gscmail.gsfc.nasa.gov...
>
>> Dick Jackson wrote:
>>
>>
>>> The ID3 (Iterative Dichotomizer - 3) method of Ross Quinlan may be what
>>> you're thinking of [...]
>>
>> However, I noticed that the web page had no links to an actual
>> implmentation.
>
>
> I just remembered that Quinlan followed up ID3 with an enhancement in 1993
> called C4.5, and published a book and C code to implement it (ISBN:
> 1558602380). If it's of any use, the code's home should be at
> http://www.cse.unsw.edu.au/~quinlan but I'm having trouble reaching it. A
> mirror and very nice tutorial are at
> http://www.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tut orial.html
>
> This code works only with text files as input and output, which might not be
> at all helpful to getting to work through IDL.

Thanks! I'll give it a try. I'm not committed to using IDL for this purpose. C code designed for a Unix platform is acceptable, and getting the data into text format won't be difficult.
If I run into any problems (or better yet, if it works!), I'll report back to you.

Subject: Re: Chi-square decision trees
Posted by Dick Jackson on Fri, 19 Apr 2002 21:36:24 GMT

"James Kuyper" <kuyper@gscmail.gsfc.nasa.gov> wrote in message
news:3CC04E6E.7060304@gscmail.gsfc.nasa.gov...
> Dick Jackson wrote:
>
>> The ID3 (Iterative Dichotomizer - 3) method of Ross Quinlan may be what
>> you're thinking of [...]
>
> However, I noticed that the web page had no links to an actual
> implmentation.

I just remembered that Quinlan followed up ID3 with an enhancement in 1993 called C4.5, and published a book and C code to implement it (ISBN: 1558602380). If it's of any use, the code's home should be at http://www.cse.unsw.edu.au/~quinlan but I'm having trouble reaching it. A mirror and very nice tutorial are at
http://www.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tut orial.html

This code works only with text files as input and output, which might not be at all helpful to getting to work through IDL.

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