Earth Surf. Dynam. Discuss.,
doi:10.5194/esurf-2016-6-RC1, 2016

Earth Surface Dynamics Discussions

Interactive comment on "An introduction to learning algorithms and potential applications in geomorphometry and earth surface dynamics" *by* Andrew Valentine and Lara Kalnins

I. Evans (Referee)

i.s.evans@durham.ac.uk

Received and published: 11 March 2016

Comments by Dr. Ian S. Evans, Durham University, 6 March 2016 on - Earth Surf. Dynam. Discuss., doi:10.5194/esurf-2016-6, 2016

An introduction to learning algorithms and potential applications in geomorphometry and earth surface dynamics. Andrew Valentine and Lara Kalnins

GENERAL COMMENTS:

This is a general survey, essentially a tutorial, of a broad field of methods. It provides a lot of common-sense advice. It is readable (more than anticipated!), and well written except for some overuse of 'we'. I am in favour of the use of personal pronouns for

emphasis, even in scientific writing, but four 'we's in four lines (p.2, lines 17-20) seems excessive. Also 'we aim to give readers. . .' (p.1 line 22) comes across as condescending: I would say 'This review is intended to provide. . .'

Figs. 2, 3, 6 and 7 use hypothetical data sets (bivariate scatters of points). This is less persuasive than the use of real examples. If made-up data have to be used, however, could they at least be more realistic? Bivariate scatters of real data tend to fade away toward the edges, and clusters are hardly ever as separated as in Figs. 6 and 7. It should not be difficult to replace these Figs. – or even to find comparable scatters that are real observations.

Many papers use K-Means as a 'black box': Fig.2 commendably provides a good illustration of how this algorithm works. The same cannot be said, however, about Fig.7 (SOM): I do not see how (b) leads on to (c), and neither the text nor the caption enlightens me.

One reason for resorting to more complex methods is because of the non-linearity of relationships (some are triangular) and the non-sphericity of clusters in property space. This can be hinted at, even in 2-D graphs: e.g. the separating line in a redrawn Fig. 6 could be curved.

The case for black boxes or learning algorithms needs to be based on the inadequacy of simpler approaches. On p.8 (line 4), porosity is considered a function of water depth and distance. That would seem to be a case for straightforward regression with two quantitative controls, on transformed scales if necessary.

The paper provides a good read and a useful overview and juxtaposition of decision trees, K-Means, PCA, Neural networks, SVM and SOM. It covers the range of statistics from Pearson to Bayes. Although I am not about to use learning algorithms, I found a few useful references. But the case made would be much more persuasive if the paper were thoroughly revised to use real examples throughout (that in Fig. 4 is appreciated); to focus on situations where use of these methods can be shown to improve results

considerably over ‘traditional’ methods; and even to show where they have led to new insights into environmental relationships.

I am relieved that the authors ‘do not advocate . . . wholesale replacement’ of ‘traditional’ approaches. My approach is to keep to the simplest methods that work, and to turn ‘black boxes’ into transparent boxes wherever possible. If a classification produces highly overlapping classes on most dimensions, I am not interested.

DETAILS:

(page/line) 2/13 in which software context should readers ‘begin by writing their own. . .’?

2/16 what are ‘optimised libraries’?

2/34 ‘as black boxes’

3/12 the example is of ‘diverse’ observations, not just multiple.

3/17 the title of section 2 led me to expect specific applications. These actually come in later sections: as 2 is very general, perhaps that should be in the heading.

4/8 to 11 this is a bit too general. Surely a classification can be evaluated only when the aspects of the data on which it is based can be understood?

5/9 why do coordinates ‘generally have little physical meaning’?

6/14 & 15 a desktop computer is contrasted with ‘a modern machine’. This is very vague. Desktops can be modern?

Fig.1 I prefer ‘standardised’ to ‘normalised’ here: the latter may imply a normal distribution.

7/3 ‘the data are perfect’

9/29 I think it is the components / new variables / dimensions that are uncorrelated – not ‘each parameter’.

Fig.3 shows an even distribution of data points within a sharply-bounded rectangle. I never saw an earth science data set like that: could we have a more realistic scatter?

Fig.4 is interesting: I see differences between the three rows. In the top row, N=100 captures the original thoroughly; in the second, N=200 is required: and in the bottom row only N=500 captures the lineations in the original. Thus I disagree with ‘using only 100-200 dimensions’, at the end of the caption.

Fig.6 Actually one dimension (y) DOES separate blue from red: they overlap only in x. This is because an unrealistically wide separation has been portrayed: again, the hypothetical example could at least approach realism.

19/6 & 7 I am unsure about ‘various’, but as it implies more than one, it should read ‘statistical distributions of the ... classifications’.

19/16 There is no point mentioning CPU hours unless some idea of the machine involved is given.

19/26 ‘as well as skill’

Interactive comment on Earth Surf. Dynam. Discuss., doi:10.5194/esurf-2016-6, 2016.