

Re: datasets to test clustering algorithm

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2004-08/0221.html>

From: Ross Clement (clemenr_at_wmin.ac.uk)

Date: 08/13/04

Date: 13 Aug 2004 00:01:14 -0700

Rajarshi Guha <rajarshi@presidency.com> wrote in message news:<pan.2004.08.12.19.58.15.292994@presidency.com>...

> *Hello, does anybody know of any freely available datasets which I could use in the comparison of a few clustering algorithms? I know of the UCI machine learning repository but I'm not sure which dataset would be useful for such a comparison. I'm looking for a large dataset (> 1000 observations) with 2 to 4 clusters.*

>

> *Any pointers would be appreciated.*

My extreme personal bias suggests the following:

Look up David Hoover's paper:

Statistical Stylistics and Authorship Attribution: an Empirical Investigation

Hoover D.L.

Literary and Linguistic Computing, december 2001, vol. 16, iss. 4, pp. 421-444(24)

Oxford University Press

This is a paper that attempts to discover the author of a paper by clustering texts based on a fixed size set of word counts in a set of documents.

You'd have to do some programming. But, if you can, you can download a whole lot of books from project gutenberg (<http://www.promo.net/pg/>), process them to turn them into these fixed-length numerical vectors (as described in more detail in his paper), and then cluster away to see if you end up clustering books by the same author together, or not.

And, if your clustering algorithms get better results than Hoover got, you could try submitting a paper on your results in that journal!

Cheers,

Ross-c

Re: datasets to test clustering algorithm