

Re: clustering question

Source: <http://sci.tech-archive.net/Archive/sci.stat.edu/2008-05/msg00021.html>

- *From:* Greg Heath <heath@xxxxxxxxxxxxxxxxxxxx>
 - *Date:* Mon, 12 May 2008 17:24:10 -0700 (PDT)
-

On May 11, 8:54 am, "ozgun.harmanci" <ozgun.harmanci@xxxxxxxxxx> wrote:

Hello,

We have been doing some data clustering to compare samples generated by two different methods: A method is used to generate sample x_1 , then we cluster x_1 using diana in R package and determine the optimal clustering scenario by maximizing calinsky harabasz index (as calculated by R). diana is divisive analysis, which is a hierarchical divisive clustering method. It computes a tree or dendrogram.

Our hypothesis is that one method should generate data which is less scattered, meaning that cluster analysis should yield less number of clusters.

However, when we do the clustering analysis on the generated samples, we saw that there is no clear distinction between number of clusters. But if I look at the tree's generated by diana then it is obvious to me that the method which we expect to have less clusters has less spread in the tree.

I am thinking that we should also use the variance of data in the clusters in addition to number of clusters to compare the sampling methods. I, however, could not find a theoretical way to do that. Could you suggest me ideas, papers or books to follow up with this problem?

I hope this makes sense.

You could define the dissimilarity between between a cluster and a point as the mahalanobis distance and ignore off diagonal terms in the covariance matrix.

Hope this helps.

Greg
w.r.t. the

.