

Re: NP-hardness of k-means clustering

Source: <http://sci.tech-archive.net/Archive/sci.math/2004-09/3537.html>

From: Eray Ozkural exa (erayo_at_bilkent.edu.tr)

Date: 09/15/04

Date: 15 Sep 2004 04:58:33 -0700

Stas Busygin <busygin@a-teleport.com> wrote in message
news:<3LK1d.24664\$zT6.4824@bignews5.bellsouth.net>...

> *Hi Eray,*

>

> *I checked the compendium first of all, and there are only two
> clustering-related problems in Misc section:*

>

> <http://www.nada.kth.se/~viggo/wwwcompendium/node126.html>

>

> *Both of them are not about k-means.*

No, they are not. This one is the closest to k-means. See comments
below.

> *Eray Ozkural exa wrote:*

> > *busygin@a-teleport.com (Stas Busygin) wrote in message
news:<8a5a72d7.0409071628.6be8cbf4@posting.google.com>...*

> >

> > > *Is there a published proof that the K-Centroid problem*

> > > *is NP-hard?*

> > >

> > > *[K-Centroid Problem]*

> > > *Given n points in R^m , find k other points in R^m*

> > > *(centroids) such that the sum of distances from each*

> > > *of the given points to its closest centroid is minimal.*

> >

> >

> > *This seems to be the more general problem. Points in R^m is a special
> > case.*

> >

> > *MINIMUM K-CENTER*

> > <http://www.nada.kth.se/~viggo/wwwcompendium/node128.html>

>

> *No, this is a different problem since it talks about "the*

> *maximum distance" and the centers are given.*

The centers are not given, only k is given, so this is indeed a very
logical and general formulation of a k-means style clustering

algorithm, in my opinion.

Read it carefully again, it was confusing to me at first, as well. If we show some effort to disambiguate the text, the "maximum distance from a vertex to its center" is being *minimized*, and hence the title MINIMUM K-CENTER. If it were being maximized, it would not be a clustering algorithm as implied in the *references* (See [1]), it would be a "distribution" algorithm instead :)

The formula after "i.e." is wrong it should have been

$$\max_{\{u \in V, c \in C\}} d(u,c)$$

And the objective is naturally

$$\min. \max_{\{u \in V, c \in C\}} d(u,c)$$

The page is in error, and I have reported this mistake.

For your specific problem in R^n , you may want to chase the references on the page.

Regards,

--

Eray Ozkural

[1] was in the bibliography of the book as:

Agarwal, P. K., and Procopiuc, C. M. (1998),

``Exact and approximation algorithms for clustering'',

Proc. 9th Ann. ACM-SIAM Symp. on Discrete Algorithms, ACM-SIAM,

658-667.