确定聚类数(Determining the Number of Clusters)

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

1.网页

Selection of K in K-means Clustering, Reloaded（gap statistic 聚类后）

https://datasciencelab.wordpress.com/2014/01/21/selection-of-k-in-k-means-clustering-reloaded/

如何确定kmeans算法的k值

https://www.jianshu.com/p/d5e61d6e746c

究竟应该聚多少类？聚类分析

https://mp.weixin.qq.com/s?__biz=MzA5NjQ3MzE2NA%3D%3D&mid=2650333123&idx=1&sn=b9514e4392126106dd39a4fe7776936f&scene=45#wechat_redirect

Deciding the Number of Clusterings

http://freemind.pluskid.org/machine-learning/deciding-the-number-of-clusterings/

K-Means算法之K值的选择

https://www.biaodianfu.com/k-means-choose-k.html

How do I determine k when using k-means clustering?

https://stackoverflow.com/questions/1793532/how-do-i-determine-k-when-using-k-means-clustering

【机器学习】确定最佳聚类数目的10种方法

https://www.cnblogs.com/think90/p/7133753.html

如何确定聚类算法中的类别个数?

https://www.zhihu.com/question/28695396

K-means怎么选K?

http://sofasofa.io/forum_main_post.php?postid=1000282

哪种聚类算法可以不需要指定聚类的个数，而且可以生成聚类的规则?

https://www.zhihu.com/question/20977382

2.论文

Hamerly G, Elkan C. **Learning the k in k-means**[C]//Advances in neural information processing systems. 2004: 281-288.

http://papers.nips.cc/paper/2526-learning-the-k-in-k-means.pdf

Pham D T, Dimov S S, Nguyen C D. **Selection of K in K-means clustering**[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2005, 219(1): 103-119.

http://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf

Ishioka T . **Extended K-means with an Efficient Estimation of the Number of Clusters**[C]// Intelligent Data Engineering and Automated Learning - IDEAL 2000, Data Mining, Financial Engineering, and Intelligent Agents, Second International Conference, Shatin, N.T. Hong Kong, China, December 13-15, 2000, Proceedings. Morgan Kaufmann Publishers Inc. 2000.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.3377&rep=rep1&type=pdf

M. J. Li, M. K. Ng, Y. Cheung and J. Z. Huang, "**Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters**," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, pp. 1519-1534, Nov. 2008, doi: 10.1109/TKDE.2008.88.

Tibshirani R, Walther G, Hastie T. **Estimating the Number of Clusters in a Data Set via the Gap Statistic**[J]. Journal of the Royal Statistical Society B, 2001, 63(2):411-423.（gap statistic 聚类后）

https://statweb.stanford.edu/~gwalther/gap

卢建云, 朱庆生, 吴全旺. **一种启发式确定聚类数方法**[J]. 小型微型计算机系统, 2018, v.39(07):7-11.

http://xwxt.sict.ac.cn/CN/article/downloadArticleFile.do?attachType=PDF&id=4503

周世兵, 徐振源, 唐旭清, et al. **新的K-均值算法最佳聚类数确定方法**[J]. 计算机工程与应用, 2010, 46(16):27-31.

http://cea.ceaj.org/CN/article/downloadArticleFile.do?attachType=PDF&id=20648

周世兵, 徐振源, 唐旭清. **K-means算法最佳聚类数确定方法**[J]. 计算机应用, 2010, 30(8):1995-1998.（聚类后）

http://kns.cnki.net//KXReader/Detail?TIMESTAMP=63706211160143500O&DBCODE=CJFQ&TABLEName=CJFD2010&FileName=JSJY201008004&RESULT=1&SIGN=wa0DAkLMK1vxouYI1Vq08jZkl3M%3d

陈黎飞, 姜青山, 王声瑞. **基于层次划分的最佳聚类数确定方法**[J]. 软件学报, 2008, 19(1):62-72.

http://www.dmi.usherb.ca/~wang/publications/journals/2008/LiFei_JoS.pdf

刘丹,高世臣.**K-均值算法聚类数的确定**[J].硅谷,2011(06):38-39.

http://kns.cnki.net//KXReader/Detail?TIMESTAMP=637062119163310000&DBCODE=CJFQ&TABLEName=CJFD2011&FileName=GGYT201106066&RESULT=1&SIGN=3BTnJ7FtnC56BeVUtHdKC8vpe9Y%3d

3.总结

1). 聚类前

Dirichlet process：采样，时间复杂度高，超参数的设定。可以借鉴。

其他聚类方法

- Canopy：需要选取T1与T2，抽样交叉验证得到。可以借鉴。
- mean-shift：需要设置核函数中的带宽大小σ。可以借鉴。
- 密度峰值聚类算法：2014年提出，需要求两两数据点之间的距离，选ρ与δ都最大的值。可以借鉴。
- AP：2007年提出，需要求两两数据点之间的距离
- 层次聚类：例如CURE算法，需要采样，时间复杂度高，设置采样参数大小、代表点数量、分区数与收缩因子。
- DBSCAN：不适合高维数据，需要设置距离半径，点数阈值。
- 。。。

2). 聚类后

聚类评价指标、准则类（AIC、BIC。。。）、Gap Statistic、轮廓系数。。。

3). 边聚类边确定聚类数

Component-Wise EM

写在最后：天下没有免费的午餐。即使消除了聚类数K这个超参，势必会引入新的超参，而这些新的超参同样对聚类的个数起了很大的影响。目前为止，并没有很好的解决方案。这也是聚类这个研究领域中一个待解决的问题。如果您遇到了好的确定聚类数的方法，欢迎交流~