CCML 2021 第十八届中国机器学习会议
The 18th China Conference on Machine Learning
2021年8月6-8日 中国·长沙
AUG 6-8, 2021 CHANGSHA, CHINA

研究背景：

传统的聚类算法，例如k-means，通过计算每个样本点与所有簇中心的距离将样本点分配给与其最近的簇。然而，异常样本点极易破坏数据的分布，致使簇中心发生较大的偏差。本文通过计算样本点与潜在簇中心的距离赋予样本点不同的权重，降低外点对数据分布的影响。此外，通过对权重向量施加 ℓ_0 范数，在聚类模型中自适应移除外点。

算法

目标函数

$$\min \sum_{i=1}^{n} s_i \sum_{j=1}^{k} \|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2^2 y_{ij} + \gamma \|\boldsymbol{s}\|_2^2.$$

$$s.t.\ \boldsymbol{Y} \in Ind, \|\boldsymbol{s}\|_0 = m, \boldsymbol{s}^T\boldsymbol{1} = 1, \boldsymbol{0} \leq \boldsymbol{s} \leq \boldsymbol{1}, C$$

求解：

固定C和s，求解Y：

$$\min_{\boldsymbol{Y} \in Ind} \sum_{i=1}^{n} \sum_{j=1}^{k} s_i \|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2^2 y_{ij}.$$

固定s和Y，求解C：

$$\sum_{i=1}^{n} \sum_{j=1}^{k} s_i (\boldsymbol{x}_i - \boldsymbol{c}_j) y_{ij} = 0$$

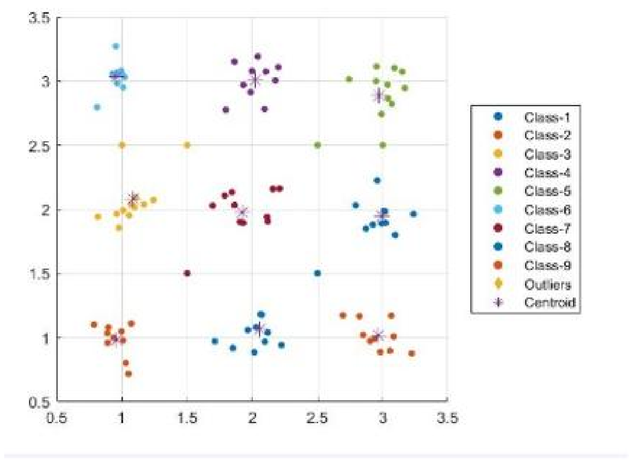
$$\Rightarrow \quad \boldsymbol{c}_j = \frac{\sum_{i=1}^{n} s_i \boldsymbol{x}_i y_{ij}}{\sum_{i=1}^{n} s_i y_{ij}}$$

固定C和Y，求解s：

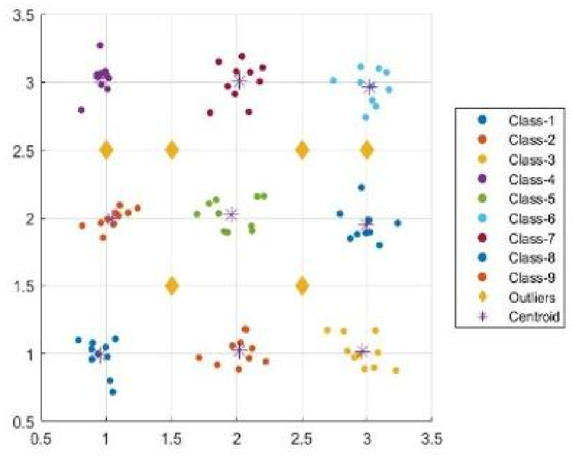
$$\min \sum_{i=1}^{n} s_i \sum_{j=1}^{k} \|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2^2 y_{ij} + \gamma \boldsymbol{s}^T \boldsymbol{s}.$$

$$s.t. \quad \boldsymbol{s}^T\boldsymbol{1} = 1, \boldsymbol{0} \leq \boldsymbol{s} \leq 1, \|\boldsymbol{s}\|_0 = m$$

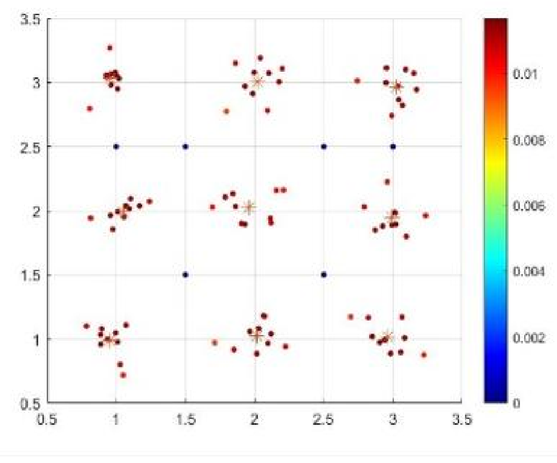
实验：



（a）k-mean 聚类结果



（b）本文算法聚类结果



（c）本文算法权重分布

数据集	NMI				Rn			
	k-means	NEO-k-means	COR	Proposed	k-means	NEO-k-means	COR	Proposed
ecoli	37.1±5.7	42.2±2.8	40.6±5.5	**42.4±3.1**	17.4±9.8	21.1±6.9	21.2±8.3	**22.4±7.8**
glass	22.9±5.3	21.6±4.4	26.3±7.8	**30.3±4.3**	11.3±4.5	9.2±3.2	**16.3±7.2**	11.1±4.2
yeast	17.0±4.6	15.9±4.9	8.9±1.5	**18.0±4.5**	8.2±4.4	5.1±4.1	1.8±1.9	**8.3±5.3**
zoo	68.1±8.2	64.3±4.4	64.6±9.1	**72.3±7.1**	54.9±14.8	54.5±9.0	59.1±11.8	**65.1±13.0**

数据集	Jaccard				F-measure			
	k-means	NEO-k-means	COR	Proposed	k-means	NEO-k-means	COR	Proposed
ecoli	2.1±5.5	26.3±13.6	4.6±11.6	**29.0±11.2**	3.7±9.4	39.5±20.4	7.0±17.4	**43.5±16.9**
glass	0.0±0.0	8.6±2.1	0.0±0.0	**32.0±46.8**	0.0±0.0	15.7±3.7	0.0±0.0	**32.0±46.8**
yeast	12.6±9.1	1.2±0.3	1.8±9.4	**23.9±40.1**	3.9±13.5	2.3±0.6	2.4±12.3	**25.6±40.8**
zoo	1.0±3.1	1.3±2.3	0.0±0.0	**5.0±21.9**	0.0±0.0	2.6±4.3	0.0±0.0	**5.0±21.9**

总结

本文提出了一种改进的k均值模型，该模型可以同时实现聚类和外点检测。为了提高模型的鲁棒性，根据数据点与潜在聚类中心的距离，对数据点分配不同的权。为此，通过对权重向量施加ℓ_0范数在聚类模型中自适应移除外点。大量的实验结果证明了该模型在聚类有效性和外点检测方面的优越性。

投稿人 姓名：胡豪杰等
单位：火箭军工程大学