TBSI-WOLT

The 3rd TBSI Workshop on Learning Theory

t-k-means: A Robust and Stable k-means Variant

Yiming Li[1], Yang Zhang[1], Qingtao Tang[1], Weipeng Huang[2], Yong Jiang[1,3], Shu-Tao Xia[1,3]

[1] Tsinghua University, [2] University College Dublin, [3] Peng Cheng Laboratory

li-ym18@mails.tsinghua.edu.cn; xiast@sz.tsinghua.edu.cn

Motivation

k-means algorithm suffers from relatively poor performance when the training set containing outliers. This drawback relates to its latent connection between the Gaussian mixture model (GMM) and the thin-tailed property of the Gaussian distribution.

Given the fact that the t-mixture model (TMM) is a robust variant of GMM, in this paper, we intend to explore the following question:

Is it possible to derive a robust variant of the k-means algorithm from the TMM, based on the relationship between the k-means algorithm and GMM?

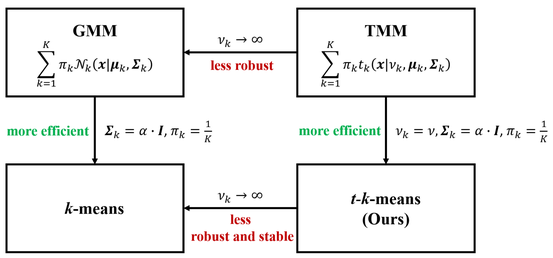

Figure 1. The relationship among our t-k-means, k-means, TMM, and GMM.

Contributions

1. We propose a novel clustering method, which is a robust and stable k-means extension.
2. We discuss the robustness and the stability of the proposed method theoretically from the aspect of the loss function and the expression of the clustering center, respectively.
3. Extensive experiments are conducted, which empirically verify the effectiveness and efficiency of the proposed method.

Useful Links

The (MATLAB) code is available at:
`https://github.com/THUYimingLi/t-k-means`

The paper is available at:
`https://arxiv.org/abs/1907.07442`

Method

1. How to Derive our t-k-means from the TMM: We assume that the data is sampled from the t-mixture distribution with the probability density function given by $t(\boldsymbol{x}|\boldsymbol{\Psi})=\sum_{k=1}^{K}\pi_k t_k(\boldsymbol{x}|\nu_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, where $\boldsymbol{\Psi}=\{\boldsymbol{\pi},\boldsymbol{\nu},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ and $\boldsymbol{\nu}=\{\nu_k|k=1,\ldots,K\}$. *The t-k-means algorithm can be derived from the TMM, under conditions that* **(1)** $\pi_k=\frac{1}{K}$, **(2)** $\boldsymbol{\Sigma}_k=\alpha\boldsymbol{I}$, and **(3)** $\nu_k=\nu$. As such, the t-k-means contains only three parameters, including the **(1)** 'deviation' α, **(2)** cluster centers $\boldsymbol{\mu}$, and **(3)** degree of freedom ν.

2. The EM-based Training Process of t-k-means: Let $\mathcal{D}=\{\boldsymbol{x}_n\}_{n=1}^{N}$ denotes the training dataset, where $\boldsymbol{x}_n\in\mathbb{R}^p$ is the p-dim sample. Suppose that the complete-data vector in TMM is indicated by $\boldsymbol{x}_c=(\boldsymbol{x}_1^\top,\ldots,\boldsymbol{x}_N^\top,\boldsymbol{z}_1^\top,\ldots,\boldsymbol{z}_N^\top,u_1,\ldots,u_N)^\top$, where $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N$ are the cluster-related indicator and $u_1,\ldots,u_N$ are the additional missing data, *s.t.*, **(1)** $\boldsymbol{x}_n|u_n,z_{nk}=1\sim\mathcal{N}(\boldsymbol{\mu}_k,\frac{\alpha\boldsymbol{I}}{u_n})$ and **(2)** $u_n|z_{nk}=1\sim\text{gamma}(\frac{1}{2}\nu,\frac{1}{2}\nu)$. The parameters $(\alpha,\boldsymbol{\mu},\nu)$ involved in the t-k-means can be iteratively estimated by the EM algorithm, as follows:

E-step (estimate the $E(z_{nk}|\boldsymbol{x}_n)$, $E(u_n|\boldsymbol{x}_n,\boldsymbol{z}_n)$, and $E(\ln u_n|\boldsymbol{x}_n,\boldsymbol{z}_n)$):

(1) $E(z_{nk}|\boldsymbol{x}_n)=\frac{t_k(\boldsymbol{x}_n|\nu,\boldsymbol{\mu}_k,\alpha\boldsymbol{I})}{\sum_{j=1}^{K}t_j(\boldsymbol{x}_n|\nu,\boldsymbol{\mu}_j,\alpha\boldsymbol{I})}\triangleq\tau_{nk}$.

(2) $E(u_n|\boldsymbol{x}_n,\boldsymbol{z}_n)=\frac{\nu+p}{\nu+\frac{1}{\alpha}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}_n-\boldsymbol{\mu}_k)}\triangleq u_{nk}$.

(3) $E(\ln u_n|\boldsymbol{x}_n,\boldsymbol{z}_n)=\ln u_{nk}+\phi\left(\frac{\nu+p}{2}\right)-\ln\left(\frac{\nu+p}{2}\right)$.

M-step (update the α, $\boldsymbol{\mu}$, and ν):

(1) $\alpha^\star=\frac{\sum_{k=1}^{K}\sum_{n=1}^{N}\tau_{nk}u_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k^\star)^\top(\boldsymbol{x}_n-\boldsymbol{\mu}_k^\star)}{p\sum_{k=1}^{K}\sum_{n=1}^{N}\tau_{nk}}$.

(2) $\boldsymbol{\mu}_k^\star=\frac{\sum_{n=1}^{N}\tau_{nk}u_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N}\tau_{nk}u_{nk}}$.

(3) $\nu^\star=\frac{1}{-\eta}$, where $\eta=1+\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\sum_{n=1}^{N}\tau_{nk}}\sum_{n=1}^{N}\tau_{nk}(\ln u_{nk}-u_{nk})+\phi\left(\frac{\nu+p}{2}\right)-\ln\left(\frac{\nu+p}{2}\right)$.

3. fast t-k-means: In TMM, if ν is unknown, the EM algorithm converges slowly. To enhance the efficiency, we fix ν as a constant as suggested in previous works. To further enhance the efficiency, we also apply $\alpha\to 0$. With fixed ν and $\alpha\to 0$, we obtain a fast version of t-k-means, which is dubbed fast t-k-means.

4. fast t-k-means++: The fast t-k-means++ is defined as a special case of the fast t-k-means, which is initialized with the result of k-means++ instead of the random initialization.

Robustness and Stability Analysis of our Method

The Robustness Analysis of t-k-means. Since the log likelihood of t-k-means is given by $\ln L(\boldsymbol{\Psi}|\boldsymbol{x})=\ln\prod_{n=1}^{N}\prod_{k=1}^{K}[t_k(\boldsymbol{x}_n|\nu,\boldsymbol{\mu}_k,\alpha\boldsymbol{I})]^{z_{nk}}$, we can rewrite its loss function and focus on the term related to the samples, as follows: $J_{t\text{-}k\text{-means}}(\boldsymbol{x},\boldsymbol{\mu})\propto\sum_{n=1}^{N}\sum_{k=1}^{K}\tau_{nk}\ln\left(1+\frac{1}{\nu\alpha}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)$. In the meanwhile, $J_{k\text{-means}}(\boldsymbol{x},\boldsymbol{\mu})\propto\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}_n-\boldsymbol{\mu}_k)$. In other words, *$J_{t\text{-}k\text{-means}}$ is a $\log\ell^2$ loss function of $\boldsymbol{x}_n$, while that of the k-means algorithm is a ℓ^2 loss*. Accordingly, t-k-means is more robust than k-means since its objective function is far less sensitive to the noise or outliers.

The Stability Analysis of t-k-means. The randomness of the k-means and t-k-means methods is mainly involved in the selection of the initial clustering centers. Once those centers are selected, the clustering results of both methods fixed. In k-means, the update of clustering center is based only on the information of samples in its cluster. In contrast, *the update of that in t-k-means is determined by the information of all samples*. In other words, no matter which sample is selected as the initial clustering center, the update of centers still depends on all samples. This use of such 'global information' significantly reduces the influence of the randomized center initialization in t-k-means algorithm, therefore it enjoys stronger stability.

Experiments

Evaluation Metric. We adopt the **(1)** adjusted rand index (ARI), **(2)** clustering mean squared error (MSE), and **(3)** W/B, to evaluate models. *The higher the ARI and the smaller the MSE and W/B, the better the method.* Among all methods, the one with the best and second-best performance is indicated in the boldface and underline, respectively.

Table 1. Results on synthetic datasets.

	A1	A2	A3	S1	S2
k-means	0.804±0.068	0.807±0.056	0.829±0.039	0.844±0.059	0.826±0.057
k-means++	0.856±0.050	0.864±0.030	0.882±0.041	0.904±0.046	0.850±0.053
k-medoids	0.775±0.081	0.783±0.057	0.792±0.039	0.817±0.056	0.803±0.070
k-medians	0.760±0.060	0.780±0.060	0.780±0.040	0.810±0.070	0.780±0.070
GMM	0.088±0.013	0.052±0.008	0.035±0.012	0.127±0.009	0.122±0.002
TMM	0.483±0.189	0.295±0.153	0.264±0.110	0.409±0.185	0.483±0.143
t-k-means	0.851±0.061	0.853±0.041	0.882±0.038	0.932±0.062	0.872±0.050
fast t-k-means	0.922±0.035	0.928±0.025	0.929±0.028	0.986±0.000	0.937±0.000
fast t-k-means++	**0.954±0.045**	**0.948±0.021**	**0.945±0.020**	**0.986±0.000**	**0.936±0.000**
	S3	S4	Unbalance	dim32	dim64
k-means	0.639±0.039	0.584±0.026	0.589±0.306	0.650±0.081	0.639±0.091
k-means++	0.671±0.035	0.589±0.028	0.909±0.078	0.985±0.028	0.995±0.018
k-medoids	0.649±0.039	0.583±0.033	0.652±0.076	0.771±0.094	0.756±0.095
k-medians	0.650±0.040	0.570±0.030	0.610±0.090	0.740±0.080	0.760±0.080
GMM	0.113±0.010	0.094±0.032	0.057±0.062	0.000±0.000	0.000±0.000
TMM	0.248±0.107	0.157±0.092	0.426±0.246	0.507±0.132	0.540±0.188
t-k-means	0.699±0.028	0.612±0.011	0.829±0.169	0.968±0.065	0.938±0.102
fast t-k-means	0.718±0.018	0.618±0.005	0.807±0.093	0.931±0.057	0.904±0.050
fast t-k-means++	**0.726±0.011**	**0.623±0.000**	**0.931±0.076**	**0.997±0.015**	**1.000±0.000**

Table 2. Results on real-world datasets.

metrics	methods	Bezdekiris	Iris	Seed	Wine
MSE	k-means	0.201±0.033	0.192±0.020	**0.420±0.000**	1.164±0.065
	k-means++	0.198±0.031	0.198±0.031	**0.420±0.000**	1.154±0.000
	k-medoids	0.205±0.037	0.222±0.048	0.426±0.003	1.278±0.159
	k-medians	0.215±0.045	0.226±0.051	0.434±0.053	1.258±0.138
	GMM	0.323±0.000	0.324±0.000	0.966±0.301	1.779±0.239
	TMM	0.425±0.238	0.293±0.106	0.606±0.121	1.612±0.209
	t-k-means	**0.186±0.000**	**0.187±0.000**	**0.420±0.000**	1.154±0.000
	fast t-k-means	0.187±0.000	**0.187±0.000**	**0.420±0.000**	**1.153±0.000**
	fast t-k-means++	0.187±0.000	**0.187±0.000**	**0.420±0.000**	**1.153±0.000**
W/B	k-means	0.225±0.050	0.213±0.030	**0.329±0.000**	0.989±0.159
	k-means++	0.222±0.046	0.222±0.046	**0.329±0.000**	**0.966±0.001**
	k-medoids	0.242±0.049	0.263±0.063	0.358±0.032	1.117±0.224
	k-medians	0.254±0.066	0.275±0.082	0.349±0.060	1.142±0.201
	GMM	0.418±0.000	0.419±0.000	2.349±3.241	3.280±2.102
	TMM	1.020±1.028	0.410±0.275	0.582±0.183	2.538±1.068
	t-k-means	**0.202±0.000**	**0.202±0.000**	0.333±0.000	1.015±0.000
	fast t-k-means	0.216±0.000	0.217±0.000	0.333±0.000	0.975±0.000
	fast t-k-means++	0.216±0.000	0.217±0.000	0.333±0.000	0.975±0.000

t-k-means variants reach better performance and smaller standard deviation than all baseline methods.

Table 3. Running cost on the Iris dataset.

Methods	Iteration	Total Time (sec)
k-means	9.58 ± 1.67	0.0159 ± 0.0048
k-means++	8.50 ± 1.81	0.0156 ± 0.0047
k-medoids	7.46 ± 1.20	0.0153 ± 0.0034
k-medians	7.64 ± 1.38	0.0186 ± 0.0056
GMM	20.00±7.84	0.1462±0.0581
TMM	28.25±8.68	0.4136±0.1299
t-k-means	29.76±5.82	0.1043±0.0228
fast t-k-means	11.78±2.05	0.0183±0.0050
fast t-k-means++	10.50±2.87	0.0181±0.0074

t-k-means is significantly faster than the TMM and its fast variants are on par with the k-means.

Acknowledgement

This work is supported by the National Science Foundation of China under Grant 61771273, the Natural Science Foundation of Zhejiang Province (LSY19A010002), the R&D Program of Shenzhen (JCYJ20180508152204044), and the Science Foundation Ireland under grant SFI/12/RC/2289_P2. The first two authors contributed equally to this work. Corresponding author: Shu-Tao Xia (email: xiast@sz.tsinghua.edu.cn).