Sample Weighting Clustering
样本加权聚类算法

Jian YU
Beijiaotong University, 100044
4/11/2006
Email: jianyu@center.njtu.edu.cn

Outline

- 1. A Brief Intro. To Cluster Analysis
- 2. Comments on weighting clustering (Nock and Nielsen,TPAMI, Aug. 2006)
- 3. How to use Maximum entropy principle to automatically determine sample weighting
- 4. Several examples and experiments
- Partial references

On cluster analysis

- With rapid development of information technology, the velocity of collecting data for human being is increasing dramatically and prior knowledge for data is not increasing at the same speed.
- Compression and Partition are two simple tools for data processing.
- Therefore, Cluster analysis are becoming popular.

Definition of cluster analysis

- partitioning a data set into most similar and meaningful subsets according to specific requirements
- What a pity，similar relation is not a equivalence relation
- Unrealistic to enumerate

How to design a clustering algorithm

- a clustering algorithm is usually developed based on a real application

According to clustering results, the existing clustering algorithms can be divided into

- Compression type: Hope to get a proper cluster prototype (C-means, FCM, EM, etc)
- Partition type: Hope to find a proper data partition(Hierarchical method, Normalized Cuts (Shi & Malik, PAMI 2000))

Compression type: C-means (MacQueen, 1967)

设 $X=\{x_1,x_2,\cdots,x_n\}\subset R^s$ 是一个数据集， $u=\{u_{ik}\}_{c\times n}\in M_{fcn}$ 是一个划分矩阵，$v=\{v_1,v_2,\cdots,v_c\}$ 是 c 个聚类中心,$v_i\in R^s$; $2\le c<n$

$$J(u,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}\left\|x_k-v_i\right\|^2$$

Where $\sum_{i=1}^{c}u_{ik}=1, u_{ik}\in\{0,1\}$

C-means

初始化:　给出初始类中心 $v^{(0)}=\left\{v_1^{(0)},v_2^{(0)},\cdots,v_c^{(0)}\right\}$, l=0,

最大迭代步数 T, 阈值　ε

Step 1: 用下列公式更新 $u_{ik}^{(l+1)}$:

$$u_{ik}^{(l+1)}=\begin{cases}1 & \mathit{if}\ i=\underset{j}{\operatorname{argmin}}\left\{\left\|x_k-v_j^{(l)}\right\|\right\}\\ 0 & \mathit{otherwise}\end{cases}$$

Step 2: 用下列公式更新 $v_i^{(l+1)}$:

$$v_i^{(l+1)}=\frac{\sum_{k=1}^{n}u_{ik}^{(l+1)}x_k}{\sum_{k=1}^{n}u_{ik}^{(l+1)}}$$

如果 $\max_i\left\|v_i^{(l+1)}-v_i^{(l)}\right\|<\varepsilon$ 或者 **l>T,** 则停止; 否则 ***l*=*l*+1** 转 **step 1.**

The advantages of C-means

- 收敛速度快,时间复杂度是线性的
- 实现简单

Drawbacks of C-means

- 1. Hard Partition
- 2. equal weight for every sample in the data set

From hard partition to soft partition

- FCM

Fuzzy C-means

- 1973, Dunn 提出了 the FCM (m=2)
- 1974, Bezdek 推广了 the FCM （ m>1)

$1 < m < +\infty$, FCM 算法的目标函数为:

$$J_m(u,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d\left(x_k, v_i\right)$$

这儿 $\sum_{i=1}^{c} u_{ik} = 1, \forall k,\quad d\left(x_k, v_i\right) = \left\| x_k - v_i \right\|^2$

FCM

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}},$$

$$u_{ik} = \frac{\left\| x_k - v_i \right\|^{\frac{-2}{m-1}}}{\sum_{j=1}^{c} \left\| x_k - v_j \right\|^{\frac{-2}{m-1}}}$$

Conditional FCM (Not equal weights but needing predefined)

- 如果对划分矩阵使用如下约束,

$$\sum_{i=1}^{c} u_{ik} = a_k \geq 0, u_{ik} \geq 0$$

其他不变,

则我们由FCM算法的目标函数可以得到Conditional FCM算法(Pedrycz, 1996)

How to determine sample weights in the literature

- Sample weights in Conditional FCM are not auto
- 文献中, 考虑样本权重的算法还有一些, 如DA(K.Rose), Weighted FCM (Karayiannis), GCM (Jian YU, 2003), 但是样本权重的确定方法一般是先验给定的

样本权重的自适应确定

- 最近,Nock and Nielsen (TPAMI, Aug. 2006)发表了一篇文章, 利用Boosting 算法的思想, 提出了一个一般的自适应样本权重聚类算法框架

Nock & Nielsen's method (1)

The data set $X = \{x_1, x_2, \cdots, x_n\}$, $\forall x_i \in R^d$, where R^d is

d-dimensional metric space, and the point x_i has the weight w_i,

l_i is the distortion function for the point x_i, then the objective funct

for the Nock and Nieslen' method can be rewritten as follows:

$$l(w) = \sum_{i=1}^{n} w_i l_i \qquad \text{where} \quad \sum_{i=1}^{n} w_i = 1, \forall w_i \geq 0 \qquad (1)$$

Nock & Nielsen's method (2)

Definition 1. The advantage over distribution w_t at iteration t is called the quantity $\gamma_t \in \Re$ that satisfies $l_{t+1}(w_t) - l_t(w) = -\gamma_t$, $\forall t \geq 0$. Vector $d_{t,i}$ is defined as $d_{t,i} = l_{t+1,i} - l_{t,i}$, $\forall 1 \leq i \leq n$.

Then they minimize a Bregman divergence to find the optimal solution as follows:

minimize $\langle 1, i_t \rangle$, where $i_t = w_{t+1.i} \ln\left(\frac{w_{t+1.i}}{w_{t.i}}\right) - w_{t+1.i} + w_{t.i}$ (2)

subject to $\langle 1, w_{t+1} \rangle = 1$, and $\langle w_{t+1}, d_t \rangle = 0$

Nock & Nielsen's method (3)

$$w_{t+1.i} = \frac{w_{t,i} \exp\left(-c_t d_{t,i}\right)}{\sum_{i=1}^{n} w_{t,i} \exp\left(-c_t d_{t,i}\right)}$$

where c_t subject to

$$\sum_{i=1}^{n} w_{t,i} d_{t,i} \exp\left(-c_t d_{t,i}\right) = 0 \quad .$$

Nock & Nielsen's method的错误

$$w_i = \begin{cases} 1 & i = \arg\min_{1 \le i \le n} l_i \\ 0 & else \end{cases}$$

是(1)和(2)的全局极小值点,即只有一点对于聚类算法为有效样本,显然这是不合理的

The Contribution of Nock & Nielsen's method

- Declaring the significance of automatically computing the sample weighting in the clustering process

How to determine sample weighting? Maximum entropy principle

- Our idea:

  Sample weighting is considered as a sampling distribution, maximum entropy principle can be applied to automatically determine sample weighting as no prior knowledge about sample weighting.

A new method to automatically compute sample weighting

- Lagrange multiplier method can result in the objective function of our new clustering algorithms as follows.

$$D = \sum_{i=1}^{n} p(x_i) l_i + \varsigma^{-1} \sum_{i=1}^{n} p(x_i) \log p(x_i)$$

where $\sum_{i=1}^{n} p(x_i) = 1, \forall p(x_i) \geq 0, \varsigma > 0$

Sample weighting equation

$$p(x_i) = \frac{\exp(-\varsigma \times l_i)}{\sum_{i=1}^{n} \exp(-\varsigma \times l_i)}$$

The impact of the parameter ζ

当 $\zeta \to \infty$, $p(x_i)$ approaches $\begin{cases} 1 & i = \arg\min_{1 \le i \le n} l_i \\ 0 & else \end{cases}$

当 $\zeta \to +0$, $p(x_i)$ approaches $\dfrac{1}{n}$

新设计的样本加权公式与原来的算法兼容

An intuitional explanation for sample weighting equation

- The larger the sample distortion from cluster prototype， the smaller the sample weighting.
- It is consistent with our intuition. Ideally, the final clustering results has no residual. Hence, the smaller sample distortion shows the importance of the sample, consistent with our equation

How to apply sample weighting equation

$$\text{Set } l_i = f\left(\sum_{j=1}^{c} \alpha_j g\left(d_{ji}\right)\right),$$

$$then\ p(x_i) = \frac{\exp\left(-\varsigma f\left(\sum_{j=1}^{c} \alpha_j g\left(d_{ji}\right)\right)\right)}{\sum_{i=1}^{n} \exp\left(-\varsigma f\left(\sum_{j=1}^{c} \alpha_j g\left(d_{ji}\right)\right)\right)}$$

GCM clustering model & its PDF

$$l_i = f\left(\sum_{j=1}^{c} \alpha_j g\left(d_{ji}\right)\right)$$

$$R = \sum_{i=1}^{n} a_i f\left(\sum_{j=1}^{c} \alpha_j g\left(d_{ji}\right)\right)$$

$$where \quad \forall j, \alpha_j \geq 0, \sum_{j=1}^{c} \alpha_j = 1$$

$$f\left(g\left(t\right)\right) = t$$

Sample weighting C-means

$$l_i = \sum_{j=1}^{c} u_{ji} \left\| x_i - v_j \right\|^2 ; \quad u_{ji} = \begin{cases} 1 & j = \arg\min_{1 \le j \le c} \left\| x_i - v_j \right\| \\ 0 & else \end{cases}$$

$$p(x_i) = \frac{\exp(-\varsigma \times l_i)}{\sum_{i=1}^{n} \exp(-\varsigma \times l_i)} ; \quad v_j = \frac{\sum_{i=1}^{n} u_{ji} p(x_i) x_i}{\sum_{i=1}^{n} u_{ji} p(x_i)}$$

Sample weighting FCM

$$l_i = \sum_{j=1}^{c} u_{ji}^m \left\| x_i - v_j \right\|^2 \; ; \quad u_{ji} = \frac{\left(\left\| x_i - v_j \right\|^2 \right)^{\frac{1}{1-m}}}{\sum_{j=1}^{c} \left(\left\| x_i - v_j \right\|^2 \right)^{\frac{1}{1-m}}}$$

$$p(x_i) = \frac{\exp(-\varsigma \times l_i)}{\sum_{i=1}^{n} \exp(-\varsigma \times l_i)} \; ; \quad v_j = \frac{\sum_{i=1}^{n} u_{ji}^m p(x_i) x_i}{\sum_{i=1}^{n} u_{ji}^m p(x_i)}$$

Sample weighting EM

$$l_i = \sum_{j=1}^{c} u_{ji} \left(\left\| x_i - v_j \right\|^2 + \beta^{-1} \ln u_{ji} \right); \quad u_{ji} = \frac{\exp\left(-\beta \left\| x_i - v_j \right\|^2 \right)}{\sum_{j=1}^{c} \exp\left(-\beta \left\| x_i - v_j \right\|^2 \right)}$$

$$p(x_i) = \frac{\exp(-\varsigma \times l_i)}{\sum_{i=1}^{n} \exp(-\varsigma \times l_i)}; \quad v_j = \frac{\sum_{i=1}^{n} u_{ji} p(x_i) x_i}{\sum_{i=1}^{n} u_{ji} p(x_i)}$$

样本加权公式的应用范围

- 容易想到，凡是可以写成（1）的聚类算法，都可以应用，实际上，公式（1）包含了大多数的压缩型聚类算法，因此，我们由样本加权公式可以得到它们的样本加权形式

On convergence of sample weighting clustering algorithms

- Sample Weighting C-means
- Sample Weighting FCM
- Sample Weighting EM

LaSalle Theorem guarantee that the above three algorithms are convergent.

On robustness of Sample weighting clustering

- 数据中加入的少量野值，对于非样本加权的聚类算法的性能影响是比较大的
- 样本加权聚类算法， 对于加入的少量野值的数据具有一定的鲁棒性。 从样本加权公式和类中心更新公式可以直观的看出这一点

实验数据

- The Iris data set has 150 data points. It is divided into three groups and two of them are overlapping. Each group has 50 data points. Each point has four attributes. More details about the IRIS data are available in Anderson [12].
- Data_3: a sample of 600 points includes 3 cluster centers:, =[0,6], =[4,0]. Each cluster consists of 200 points and the points in the ith cluster obey the normal distribution.

加入野值

Table 1. Outlier samples for two data sets

Data set	IRIS	Data_3
Added outlier samples	[100,100,100,100]	[9,-8]

实验结果 (1)

Table 2. Average and Minimum Error number of clustering results for C-means, FCM and EM in 100 runs

	IRIS	IRIS*	Data_3	Data_3*
C-means	(44.17,16)	(56,50)	(4.8,0)	(8.2,0)
FCM(m=2)	(16,16)	(50,50)	(0,0)	(0,0)
DA(β =2)	(16.68,16)	(50,50)	(0.2,0)	(1.2,0)

实验结果 (2)

Table 3. Average number and of clustering results for WCM, WFCM and WEM in 100 runs

$\varsigma=$	0.0001	0.0002	0.0003	0.0006	0.0011	0.0021	0.004	0.007	0.0127	0.0234	0.0428	0.0785	0.1438	0.2637	0.4833	0.8859	1.6238	2.9764	5.5	10
	Sample Weighting C-means																			
IRIS	32.38	36.36	36.38	45.68	42.03	44.03	36.71	43.01	38.01	38.39	44.86	44.37	34.76	37.58	41.69	41.91	52.88	53.51	54	56.27
IRIS*	26.31	25.86	26.66	24.98	26.48	26.17	22.2	22.64	22.85	27.02	23.63	26.16	26.97	27.81	44.75	52.43	50.72	49.76	50	52.27
Data_3	7.2	4.4	4.8	4.8	4.6	6.8	7.2	6.6	3.6	4.8	5.4	5.57	7.55	5.12	7.37	7.49	6.5	7.39	7.25	8.53
Data_3*	4.4	4.6	6.2	3.8	4	4.6	6.8	6.2	9	5.8	5.13	5.98	8.33	9.57	7.54	9.3	10.65	7.77	7.4	4.4
	Sample Weighting Fuzzy C-means (m=2)																			
IRIS	16	16	16	16	16	16	16	16	16	16	16	16	15	15	15	41	42	43	44	34.88
IRIS*	23	17	16	16	16	16	16	16	16	16	16	16	15	15	15	41	42	43	44	33.33
Data_3	0	0	0	0	0	0	0	0	0	0	0	0	0	7.43	11.67	19.78	24.67	24	10	5
Data_3*	0	0	0	0	0	0	0	0	0	0	0	0	0	2.57	10.69	18.91	22.28	23.15	12	5.84
	Sample Weighting EM Clustering (β=2)																			
IRIS	17.36	18.04	16.68	16.68	16	16.34	16.34	16.34	16	17.02	15	15	15	15	35.87	47	51.36	51	51	49.92
IRIS*	17	16	16	16	16	16	16	16	16	16	15	15	15	15	41	47	51.02	51	50.28	57.28
Data_3	0.4	0.4	0.6	1	1	0.2	0.2	0.4	0.4	0.6	0.6	0.76	1.07	2.88	10.37	15.03	10.32	29.89	20.22	23.27
Data_3*	0.4	1	0.6	0.4	0.6	0.2	0.8	1	2	0.8	0.8	1.52	0.77	1.62	5.03	13.98	10.92	31.7	19.06	25.15

Outlook 1： Parameter selection

- How to select a proper ζ
- For sample weighting EM, $\zeta < \beta$ should hold.
- For other sample weighting clustering algorithms, it is open.

Outlook 2. Outlier detection

- Intuitionally, it is possible to detect outlier point by sample weighting method.
- How to realize this point

Outlook 3 Switch regression

- Since clustering algorithms can be used in the regression problem, sample weighting equation can be easily applied to switch regression problem

References

- MacQueen J., Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist, Prob., (1967)1:281-297.
- Bezdek, J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York,1981
- A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” J. Royal Statistical Soc. B, vol. 39, pp. 1-38, 1977.
- W. Pedrycz, “Conditional fuzzy c-means, “ Pattern Recognition Letters, vol. 17, pp. 625 –632, 1996
- K. Rose, “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems”， Proceedings of the IEEE, 86(11), 2210 –2239, 1998
- Jian Yu, “General c-means clustering model ”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1197-1211, Aug. 2005
- Richard Nock and Frank Nielsen, “On weighting exponent”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, NO. 8, (2006) , 1223-1235.
- Jian Yu, “Analysis of weighting exponent ”, Technical report of Institute of Computer Science, Beijing Jiaotong University, 2006
- Bezdek, J.C., R.J. Hathaway, M.J. Sabin and W.T. Tucker.Convergence theory for fuzzy c-means: counterexamples and repairs. IEEE Trans. Syst. Man Cybernet. 17, 873-877. 1987
- R.N.Dave &R. Krishnapuram, “Robust clustering methods: a unified view”, IEEE Trans. Fuzzy Systems, 5(2), 270-293, May 1997.
- W. I. Zangwill, Nonlinear Programming: A Unified Approach. Englewood Cliffs, NJ: Prentice-Hall, 1969.
- Anderson E., “The IRISes of the Gaspe Peninsula,” Bull. Am. IRIS Soc., (1935) vol. 59, 2-5.
- Jian Yu, “sample weighting clustering”, Technical report of Institute of Computer Science, Beijing Jiaotong University, 2006
- Jian Yu, Yan Zhu, “sample weighting regression”, Technical report of Institute of Computer Science, Beijing Jiaotong University, 2006

Thanks for your attention!

☺