Computer Science and Application 计算机科学与应用, 2017, 7(10), 994-1001
Published Online October 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.710112

Hans汉斯

An Entropy Clustering Method for the Model and Its Algorithm of the Maximizing a Submodular Function Subject to a Matroid Constraint

Guohong Liang[1,2], Ying Li[1], Meng Ye[3], Bingjie Li[2]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an Shaanxi
[2]School of Science, Air Force Engineering University, Xi'an Shaanxi
[3]The 94826th Military Forces, Shanghai
Email: liangguohong321@163.com

Received: Oct. 6th, 2017; accepted: Oct. 19th, 2017; published: Oct. 24th, 2017

Abstract

This paper proposes a new clustering objective function with information entropy, which is composed of entropy rate of random path based on graph theory and balance item. Entropy rate is conducive to compact and uniform clustering, the balance function encourages objects with high similarity to cluster, and punishes those objects with low similarity. First, the weighted undirected graph associated with data is constructed, and it is found that this structure induces a matroid, a combination of the structure of linear independent concept in vector space. Then, the model of which is maximizing a submodular function under the constraints of the matroid is obtained. Finally, according to the monotonicity, increment and submodular of the objective function, an efficient greedy algorithm is developed and its performance guarantee is discussed.

Keywords

Clustering, Graph Theory, Information Theory, Submodular Function, Discrete Optimization

拟阵约束下最大化子模函数的模型及其算法的一种熵聚类方法

梁国宏[1,2]，李　映[1]，叶　萌[3]，李炳杰[2]

[1]西北工业大学计算机学院，陕西　西安
[2]空军工程大学理学院，陕西　西安
[3]94826部队，上海

文章引用: 梁国宏，李映，叶萌，李炳杰. 拟阵约束下最大化子模函数的模型及其算法的一种熵聚类方法[J]. 计算机科学与应用, 2017, 7(10): 994-1001. DOI: 10.12677/csa.2017.710112

Email: liangguohong321@163.com

收稿日期：2017年10月6日；录用日期：2017年10月19日；发布日期：2017年10月24日

摘　要

本文提出了一个新的带有信息熵的聚类目标函数，它是由基于图论的随机路径的熵率和平衡项两部分组成。熵率有利于形成紧凑和均匀的聚类，平衡函数鼓励相似度比较高的对象才能聚类，并惩罚那些相似度比较低的对象。首先构造了与数据关联的赋权无向图，并发现这种构造诱导出一个拟阵，它是一个组合在向量空间中推广线性独立概念的结构。接着得到了拟阵约束下最大化子模函数的模型。最后根据目标函数的单调性、递增性和下模性，开发了一个高效的贪婪算法并讨论了它的性能保证。最后根据数值实验，与已有的算法做了比较，说明了该算法的有效性。

关键词

聚类，图理论，信息理论，子模函数，离散优化

1. 引言

聚类是在机器学习中的一种无监督学习方式，既是数据挖掘的重要部分，又是模式识别领域的基础问题[1] [2]，在统计学里也有广泛的引用。在几乎每个处理经验数据的科学领域里，研究人员试图通过识别相似字符组的数据获得第一印象。在不同的领域已经提出了很多种聚类方法，并且都有性能保证。然而，它们通常是基于不同的假设，而且很难在同一个标准下比较。此外，最理想的标准会导致成 NP-hard 问题。因此聚类的进一步发展就是对存在理论证明或更新的问题的目标函数的精细化设计。

在各种各样的聚类算法中，一些使用单个目标函数，一些递归地使用中间成本函数，以及一些基于数据点的投影(子空间，流形)。制定聚类问题作为一个图拓扑选择问题，当数据点及其对应关系分别映射到图中的顶点和边。然后通过查找图拓扑来解决聚类问题。主要考虑紧凑的、均匀的、平衡的类别。在一个紧凑的类别中，数据点彼此接近。为了获得以上这些性质，我们提出了一个新的由两部分组成的目标函数：一是图上的随机路径的熵率；二是类分布的平衡项。熵率[3]有利于形成紧凑和均匀的聚类，平衡函数鼓励相似度比较高的聚类，并惩罚那些相似度比较差的对象。根据图中的随机路径和类别的分布有很大的不确定性，构造一个图拓扑。

本文把聚类作为一个图划分问题，将图划分为 k 个群，研究具有 k 连通子图的图拓扑，并最大化所提出的目标函数。

2. 模型

2.1. 基础知识

图：无向图表示为 $G=(V,E)$，其中 V 是顶点的集合，E 是边的集合。v_i 和 $e_{i,j}$ 表示顶点与边。$w_{i,j}$ 表

示顶点 v_i 和 v_j 之间边 $e_{i,j}$ 上的权重。在无向图中边的权重是对称的，即 $w_{i,j}=w_{j,i}$。

图的划分[4]：在 $G=(V,E)$ 中，如果 $V=\{V_1,V_2,\cdots,V_K\}$，且满足：

(1) $V_i\cap V_j=\Phi, i\neq j$；(2) $\bigcup_{i=1}^{K}V_i=V$，

则称 $V_1,V_2,\cdots,V_K$ 为 V 的一个划分。图的顶点的子集的选择问题就是图的划分问题。

本文的目标是选择边的子集 $A\in E$，使得子图 (V,A) 是 K-连通的。

熵：用来度量随机变量的不确定性。设离散随机变量 X，概率密度函数为 p_X，则它的熵定义为：

$$H(X)=-\sum p_X(x)\log p_X(x), \tag{1}$$

相应的，条件熵 $H(X|Y)$ 的定义为：

$$\begin{aligned}H(X|Y)&=-\sum p_Y(y)H(X|Y=y)\\&=-\sum p_Y(y)\sum p_{X|Y}(x|y)\log p_{X|Y}(x|y),\end{aligned} \tag{2}$$

其中 $p_{X|Y}$ 是条件概率密度函数[5] [6]。

熵率：用来度量随机过程 $X=\{X_t|t\in T\}$ 的不确定性。对于一个离散的随机过程，熵率定义为一个渐近测度[7]：

$$H(X)=\lim_{t\to\infty}H(X_t|X_{t-1},X_{t-2},\cdots,X_1), \tag{3}$$

对于一阶平稳的马尔可夫过程，熵率有一个简单的形式[8] [9]：

$$H(X)=\lim_{t\to\infty}H(X_t|X_{t-1})=\lim_{t\to\infty}H(X_2|X_1)=H(X_2|X_1). \tag{4}$$

第一个等式是由于一阶马尔可夫性，而第二个等式是平稳性的结果。有关更多详细信息，可以参考[10]。

图上的随机路径：设 $X=\{X_t|t\in T,X_t\in V\}$ 是带有非负的相似权 ω 的图 $G=(V,E)$ 上的随机路径。转移概率定义为：

$$p_{i,j}=\mathrm{P}(X_{t+1}=v_j|X_t=v_i)=\frac{\omega_{i,j}}{\omega_i}, \tag{5}$$

其中 $\omega_i=\sum_{k:e_{i,k}\in E}\omega_{i,k}$ 是顶点 v_i 发生事件之和，并且平稳分布为：

$$\mu=(\mu_1,\mu_2,\cdots,\mu_{|V|})^{\mathrm{T}}=\left(\frac{\omega_1}{\omega_T},\frac{\omega_2}{\omega_T},\cdots,\frac{\omega_{|V|}}{\omega_T}\right)^{\mathrm{T}}, \tag{6}$$

其中 $\omega_T=\sum_{i=1}^{|V|}\omega_i$ 是归一化常数。对于一个非连通图，平稳分布不是唯一的。但是，(6)式中的 μ 总是平稳分布，容易证明得到 $\mu=P^{\mathrm{T}}\mu$，其中 $P=[p]_{i,j}$ 是转移矩阵。随机路径的熵率为：

$$\begin{aligned}H(X)&=H(X_2|X_1)=\sum_i\mu_iH(X_2|X_1=v_i)\\&=-\sum_i\sum_j\frac{\omega_{i,j}}{\omega_T}\log\frac{\omega_{i,j}}{\omega_T}+\sum_i\frac{\omega_i}{\omega_T}\log\frac{\omega_i}{\omega_T}\end{aligned} \tag{7}$$

下模性[11]：设 E 是有限集，函数 $f:2^E\to R$ 是下模函数当且仅当 $\forall A,B\subseteq E,\forall a_1,a_2\in E$，有

$$f(A)+f(B)\geq f(A\cup B)+f(A\cap B).$$

或 $$f(A\cup\{a_1\})-f(A)\geq f(A\cup\{a_1,a_2\})-f(A\cup\{a_2\}) \tag{8}$$

等价于：$\delta f_{a_1}(A) \geq \delta f_{a_1}(A \cup \{a_2\})$，对于所有的 $A \subseteq E$ 和 $a_1, a_2 \in E \setminus A$，其中

$$\delta f_{a_1}(A) = f(A \cup \{a_1\}) - f(A)$$

进一步地，如果 $f(A) \leq f(A \cup \{a_1\})$，则称 f 是单调递增函数[12]。

拟阵：设 E 是有限集，I 是 E 的子集组成的集合，拟阵是一个有序对 $M = (E, I)$，并且满足：

1) $\Phi \in I$。
2) 如果 $I_1 \in I$ 且 $I' \subseteq I_1$，则 $I' \in I$。
3) 如果 $I_1 \in I, I_2 \in I$ 且 $|I_1| < |I_2|$，则存在元素 $e \in I_2 - I_1$，使得 $I_1 \cup e \in I$。

我们提出聚类作为一个图的划分问题。将图划分为 k 个类，搜索图具有 k 连通子图的拓扑，然后最大化提出的目标函数。

2.2. 图的构造

将数据集映射到图 $G = (V, E)$，顶点表示该数据点和边权重表示两数据点之间的相似性。为了聚类，将数据集映射成 K-最近邻域图。目标是划分图成为几个连通分支。选择一个边的子集 $A \subseteq E$，产生一个子图 $G' = (V, A)$，包含 K-连通子图。再者，我们也假设每个顶点具有自环。虽然自环不影响图的划分，但它们对提出的随机路径模型是必要的。当某边不包含在 A 内时，增加与该边关联顶点的自环的权重，这样每个顶点的总发生的权保持常数。

如果边 $e_{i,j}$ 是被聚类时选择了，则顶点 v_i 的自环的权为 $\omega_{i,i}$，顶点 v_j 的自环的权为 $\omega_{j,j}$，边 $e_{i,j}$ 的权为 $\omega_{i,j}$。

如果边 $e_{i,j}$ 是被聚类时未选择，则顶点 v_i 的自环的权为 $\omega_{i,i} \to \omega_{i,i} + \omega_{i,j}$，顶点 v_j 的自环的权为 $\omega_{j,j} \leftarrow \omega_{j,j} + \omega_{i,j}$，边 $e_{i,j}$ 的权为 $\omega_{i,j} \leftarrow 0$。

2.3. 平衡函数

利用平衡函数鼓励大小相似的数据点分到同一个聚类中。设 A 是已经被选择的边集，N_A 是图中被划分部分的数量，Z_A 是聚类的分布。例如，设图对边集 A 的划分是 $S_A = \{S_1, S_2, \cdots, S_{N_A}\}$，有

$$p_{Z_A}(i) = \frac{|S_i|}{|V|}, \quad i = \{1, 2, \cdots, N_A\}, \tag{9}$$

并且平衡项为：

$$\mathrm{B}(A) \equiv H(Z_A) - N_A = -\sum_i p_{Z_A}(i) \log\left(p_{Z_A}(i)\right) - N_A. \tag{10}$$

熵 $H(Z_A)$ 将大小相似的聚为一类，同时通过类与类数据点之间的最小化可以得到聚类数目 N_A。

推论 2： 2.2 中构造的图上的平衡函数 $\mathrm{B}: 2^E \to R$ 是一个单调递增的下模函数。

2.4. 聚类函数

目标函数是熵率和平衡函数的结合，因此得到了更加紧凑的、均匀的和平衡的聚类。

$$\begin{aligned} &\max_{A \subseteq E} \quad F(A) \\ &\text{s.t.} \quad N_A \geq K \end{aligned} \tag{11}$$

其中 $F(A) = H(A) + \lambda \mathrm{B}(A)$ 是目标函数。参数 $\lambda \geq 0$ 是平衡项的权重。目标函数也是单调递增的下模函数。

推论 3： 设 E 是边集，I 是边集的集合，$A \subseteq E$ 满足：1) A 是无循环的；2) A 构成一个大于或等于

K-连通分支的图划分，则 $M=(E,I)$ 是一个拟阵。

在无循环的约束下，图的划分问题变成了在拟阵约束下子模函数的最大化问题，即：

$$\begin{aligned}&\max_{A\subseteq E} \quad F(A)\\ &\text{s.t.} \quad A\in I\end{aligned} \tag{12}$$

3. 算法

对于求解模型(12)的贪婪算法如下：

贪婪算法：目标函数定义为 $F(A)=H(A)+\lambda \mathrm{B}(A)$

初始条件：$G=(V,E),\omega:E\to R^{+},K,\lambda$

输出：A

$A\leftarrow\varnothing, U\leftarrow E$

循环

$$\hat{a}\leftarrow\arg\max_{a\in U} \quad F(A\cup\{a\})-F(A)$$

如果 $A\cup\{\hat{a}\}\in I$，则

$A\leftarrow A\cup\{\hat{a}\}$

$U\leftarrow U-\{\hat{a}\}$

直到 $U=\varnothing$.

拟阵约束的下模函数的最大化问题是组合优化研究中活跃的领域[13]，Fisher [14]等给出了单调递增下模函数的最大化问题的以 $\frac{1}{2}$ 近似上界的贪婪算法。同样地，我们也给出如下的性能保证：

定理：设 A_{opt} 是问题(15)的最优解，A_{greedy} 是应用上述算法得到的一个近似解，则

$$\frac{F(A_{greedy})-F(\varnothing)}{F(A_{opt})-F(\varnothing)}\geq\frac{1}{2}$$

证明利用[14] theorem2.1 直接得到。

4. 实验

在聚类方面我们进行了大量的实验来评估所提出的算法方法，在整个实验中，使用了 $\lambda'=0.5$ 来测定平衡权重。该算法需要成对相似性数作为输入，这里使用的是高斯核 $\omega(v_i,v_j)=\exp\left(-\frac{d(v_i,v_j)^2}{2\sigma^2}\right)$，这里 $d(v_i,v_j)$ 是样本 i 和 j 之间的距离，σ 是核带宽。然后构造一个邻域图 1，其中在群集之前，每个示例都连接到其 30 个最近的邻域。

在实验中，我们设置了聚类的个数对于所有算法的数字 K，为了比较，使用以下两个标准的群集性能指标：1) 聚类精度(ca)和 2) rand 指数(ri)：

聚类精度是一种分类精度性能指标。设 $\mathbb{C}=\{C_1,C_2,\cdots,C_K\}$ 是聚类的真分布。类似地，设 $S=\{S_1,S_2,\cdots,S_K\}$ 是计算的聚类的分布。聚类精度为：

$$CA=\max_J\frac{1}{n}\sum_i\left|C_i\cap S_{J(i)}\right|,$$

其中 n 是数据中的样本总数和 J 表示任意可能的排列 $\{1,2,\cdots,K\}$。

聚类指数是两者之间的相似性度量。设 TP 是相同的样本对的数量群集的真实性和估计聚类，设 TN 是在中的样本对的数目真实与估计的不同聚类聚类，设 FP 是对的示例对的数量在不同的集群中为真实的集群在同一个组中进行估计聚类，FN 是样本对的数目这在同一个群中是真实的聚类，但在不同的集群中估计的聚类输出。聚类指数为：

$$RI = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

我们比较我们的结果和已有的聚类算法包括 AP、k-means、ncut，以及平面最大间隔聚类算法(cpmmc) [16]，它们代表了各种各样的聚类，结果如下：

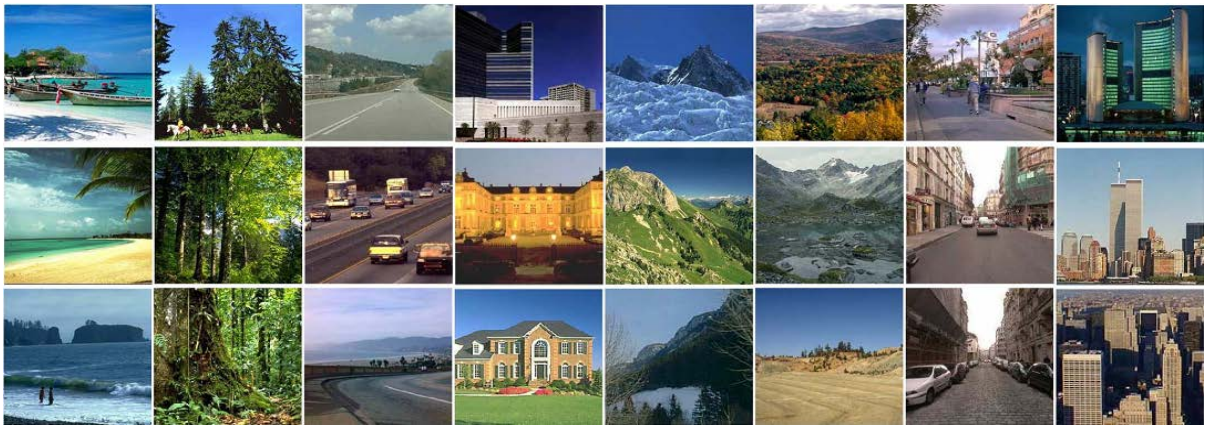

Figure 1. The image comes from the natural scene recognition dataset [15]. From left to right, the image is coast, forest, highway, inner city, mountains, open country, streets and tall buildings. Because of different imaging conditions, the same kind of image shows a great change place and season

图 1. 图像来自自然场景识别数据集[15]。从左到右，图像类是海岸、森林、高速公路、内部城市，山，开放的乡村，街道和高楼。由于不同的成像条件，相同类的图像表现出很大的变化地点和季节

Table 1. Clustering performance comparison: clustering accuracy

表 1. 聚类性能比较：聚类精度

数据	新方法	ncut	AP	k-means	cpmmc
Ionosphere	92.54	83.19	70.94	70.00	75.48
Letters	94.44	94.28	91.83	93.38	95.02
Sattellite	98.50	97.50	62.30	94.10	98.79
Digits 0689	97.34	91.83	90.31	78.46	96.74
Digits 1279	98.23	91.70	85.51	89.32	94.52
Breast Cancers	95.78	92.09	93.32	91.04	n/a
Iris	93.01	86.67	86.00	83.33	n/a
Wine	96.63	98.31	93.82	96.63	n/a
Glass	50.98	55.41	40.19	45.33	n/a
Movement Libras	52.98	50.83	46.94	44.44	n/a
Natural Scenes	47.45	56.36	43.64	47.70	n/a
MPEG-7 Shapes	73.56	71.64	69.14	n/a	n/a

Table 2. Clustering performance comparison: clustering index
表 2. 聚类性能比较：聚类指数

数据	新方法	ncut	AP	k-means	cpmmc
Ionosphere	0.87	0.72	0.59	0.58	0.65
Letters	0.89	0.89	0.85	0.88	0.92
Sattellite	0.98	0.95	0.53	0.89	0.97
Digits 0689	0.99	0.93	0.92	0.87	0.97
Digits 1279	0.95	0.92	0.87	0.90	0.96
Breast Cancers	0.86	0.85	0.88	0.84	n/a
Iris	0.92	0.86	0.85	0.83	n/a
Wine	0.97	0.98	0.92	0.95	n/a
Glass	0.72	0.70	0.66	0.70	n/a
Movement Libras	0.91	0.92	0.91	0.91	n/a
Natural Scenes	0.80	0.84	0.81	0.83	n/a
MPEG-7 Shapes	0.98	0.99	0.99	n/a	n/a

从表 1 和表 2，我们看到了所提出的算法在聚类中产生略好的性能。根据聚类精度度量，在 12 个数据集中的 7 个数据集上的算法优于其它算法。我们还获得更好的性能索引：在 12 个数据集中有 8 个更好。

基金项目

陕西省自然科学基金资助项目(20125153025)；中国高等教育博士生研究基金资助项目(20126102110041)。

参考文献 (References)

[1] Liu, M.Y., Tuzel, O., Ramalingam, S. and Chellappa, R. (2014) Entropy-Rate Clustering: Cluster Analysis via Maximizing a Submodular Function Subject to a Matroid Constraint. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[2] Ye, X. and Guo, L.J. (2012) Consructing Affinity Matrix in Spectral Clustering Based on Neighbor Propagation. *Neurocomputing*.

[3] Lin, H. and Bilmes, J. (2011) Word Alignment via Submodular Maximization over Matroids. *Proceedings of the* 49*th Annual Meeting of the Association Computational Linguistics*: *Human Language Technologies—Short Papers*, **2**, 170-175.

[4] Cao, J., Chen, P., Zheng, Y. and Dai, Q. (2013) A Max-Flow-Based Similarity Measure for Spectral Clustering. *ETRI Journal*, **35**.

[5] Chakrabarti, A. and Kale, S. (2013) Submodular Maximization Meets Streaming: Matchings, Matroids, and More. *Data Structures and Algorithms*.

[6] Liu, M.Y., Tuzel, O., Ramalingam, S. and Chellappa, R. (2011) Entropy Rate Superpixel Segmentation. *IEEE Conference on Compute Vision and Pattern Recognition.*

[7] Zhang, X., Li, J. and Yu, H. (2011) Local Density Adaptive Similarity Measurement for Spectral Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**, 352-358.

[8] Cover, T.M. and Thomas, J.A. (1991) Elements of Information Theory. 2nd Edition, John Wiley & Sons. https://doi.org/10.1002/0471200611

[9] Hua, L. (2014) Application of Spectral Clustering and Entropy in Clustering. Zhejiang University, 25-29.

[10] Liu, M.-Y., Tuzel, O., Ramalingam, S. and Chellappa, R. (2014) Entropy-Rate Clustering: Cluster Analysis via Maximizing a Submodular Function Subject to a Matroid Constraint. *Pattern Analysis and Machine Intelligence*, **36**, 99-105.

[11] Nemhauser, G.L., Wolsey, L.A. and Fisher, M.L. (1978) An Analysis of the Approximations for Maximizing Submodular Set Functions. *Mathematical Programming*, **14**, 265-294. https://doi.org/10.1007/BF01588971

[12] Oxley, J. (1992) Matroid Theory. Oxford Univ. Press, Oxford.

[13] Badanidoyuri, A. and Vondrak, J. (2014) Fast Algorithms for Maximizing Submodular Functions. *Proceedings of the* 25*th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1497-1514. https://doi.org/10.1137/1.9781611973402.110

[14] Fisher, M.L., Nemhauser, G.L. and Wolsey, L.A. (1978) An Analysis of the Approximations for Maximizing Submodular Set Functions. *Mathematical Programming*, **8**, 73-87. https://doi.org/10.1007/BFb0121195

[15] Oliva, A. and Torralba, A. (2001) Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, **42**, 145-175. https://doi.org/10.1023/A:1011139631724

[16] Frey, B.J. and Dueck, D. (2007) Clustering by Passing Messages between Data Points. *Science*, **315**, 972-976. https://doi.org/10.1126/science.1136800