机器学习中谱聚类方法的研究*)

高 琰 谷士文 唐 琎 蔡自兴
(中南大学信息科学与工程学院 长沙 410075)

摘 要 最近几年,谱聚类方法在模式识别中得到了广泛的应用。与传统的聚类方法比较,它具有能在任意形状的样本空间上聚类,且收敛于全局最优解的优点。本文着重介绍了谱方法的基本原理、相应的算法、研究状况及其在模式识别领域中的应用,同时指出了它的关键问题与未来的研究方向。

关键词 谱聚类,机器学习,图划分

Research on Spectral Clustering in Machine Learning

GAO Yan GU Shi-Wen TANG Jin CAI Zi-Xing
(College of Information Science and Engineering, Central South University, Changsha 410075)

Abstract Recently spectral clustering has wide application in pattern recognition. Comparing with traditional clustering methods, it has "global"optima solution. This paper introduces the principle, algorithms of spectral clustering, and the application in pattern recognition. And also it points out the key problems and future direction.

Keywords Spectral clustering, Machine learning, Graph partition

1 引言

聚类分析是机器学习的经典问题。它是通过抽取数据的"潜在"结构,将相似数据组成类或类的层次结构。它不需要先验知识和假设,故它也称作是无监督学习。传统的聚类算法主要是 k-means 算法和 EM 算法。这些算法都是建立在凸球形的样本空间上。当样本空间不为凸时,算法会陷入"局部"最优。

为了能在任意形状的样本空间上聚类,且收敛于全局最优解,研究学者最近开始利用谱方法来聚类。谱方法聚类是由数据点间相似关系建立矩阵,获取该矩阵的前 n 个特征向量,并且用它们来聚类不同的数据点。谱聚类方法建立在图论中的谱图理论上。最初,它是用于负载均衡和并行计算,VLSI 等方面,如 Hagen 和 Kahng[1] 将基于 ratio-cut 的目标函数图划分算法用于 VLSI 设计中。最近,学者们也开始将谱聚类方法用于机器学习中。Shi 和 Malik[2] 在 2000 年根据谱图理论建立了 2-way 划分的 Normailized-Cut(Ncut)的目标函数,设计了用于图像分割的算法,由此发展出一系列算法:k-way 划分的 Ncut 算法(Ng 和 Weiss[3]);Normalized Cut 与随机游动关系的算法(Meila 和 Shi[4]);基于二分图的算法(Zha[5]和 Dhillon[6])等。并且,谱聚类方法的应用也开始从图像分割领域扩展到文本挖掘(Dhillon[6])和生物信息挖掘领域(Chris Ding[7])等领域中。

本文着重介绍了谱方法的基本原理、相应的算法、研究状况及其在模式识别领域中的应用,同时指出了它的关键问题与未来的研究方向。

2 基本原理与算法描述

2.1 图划分问题与聚类

聚类算法的一般原则是:类内样本间的相似度大,类间样本间的相似度小。假定将每个数据样本看作图中的顶点 V,根据样本间的相似度将顶点间的边赋权重值,就得到一个基于样本相似度的无向加权图:$G(V,E)$。那么在图 G 中,我们可将聚类问题转变为如何在图 G 上的图划分问题。划分的原则是:子图内的连权重最大化和各子图间的边权重最小化。

针对这个问题,Shi 和 Malik[2] 提出了基于将图划分为两个子图的 2-way 目标函数 Ncut:

$$\min Ncut(A,B)=\frac{cut(A,B)}{vol(A)}+\frac{cut(A,B)}{vol(B)} \tag{1}$$

$$vol(A)=\sum_{i\in A}\sum_{j\sim}w_{ij} \tag{2}$$

$$cut(A,B)=\sum_{i\in A,j\in B}w_{ij} \tag{3}$$

其中 $cut(A,B)$ 是子图 A,B 间的边,又叫"边切集"。

从式(1),我们可以看出改进后目标函数不仅满足类间样本间的相似度小,也满足类内样本间的相似度大。

$$asso(A)=\sum_{i\in A}\sum_{j\in A,i\sim j}w_{ij}=vol(A)-cut(A,B) \tag{4}$$

$$\min Ncut(A,B)=\min(2-\frac{asso(A)}{vol(A)}+\frac{asso(B)}{vol(B)}) \tag{5}$$

如果考虑同时划分几个子图的话,则基于 k-way 的 Normalized-cut 目标函数为:

$$Ncut(V_1,\cdots,V_k)=\frac{cut(V_1,V_1^c)}{\sum_{i\in V_1}\sum_j W_{ij}}+\cdots+\frac{cut(V_k,V_k^c)}{\sum_{i\in V_k}\sum_j W_{ij}} \tag{6}$$

除了 Ncut 目标函数外,还有 Hagen 和 Kahng 提出的 RatioCut[1] 和 Ding 等提出的 MinMaxCut[8]。三个目标函数中,RatioCut 只考虑类间相似性最小,且最易产生"倾斜"的划分。而 MinMaxCut 与 Ncut 一样满足类内样本间的相似度大而类间样本的相似度小的原则,与 Ncut 具有相似的行为。

2.2 谱图理论

谱图理论是一个具有很长历史的理论。它是利用矩阵理论和线性代数理论来研究图的邻接矩阵,根据矩阵的谱来确定图的某些性质。谱图理论分析的基础是图的 Laplacian 矩阵,它是 Fiedler[9] 1973 提出来的。

假设一无向加权图 $G=\langle V,E\rangle$,其表示形式为一对称矩阵:$W=[W_{ij}]_{n\times n}$,其中 W_{ij} 表示连接顶点 i 与 j 的权值。那么该图的 Laplacian 矩阵表示为:

*)基金项目:国家自然科学基金重点项目(60234030)。高 琰 博士生,研究方向:智能信息处理;谷士文 教授,博导;唐 琎 副教授;蔡自兴 博导。

$$L=D-W \tag{7}$$

其中，D 为对角阵，$D_{ii}=\sum_{i\sim j}w_{ij}$。

Laplacian 矩阵是对称半正定矩阵，因此它的所有特征值是实数且是非负的：

$0=\lambda_1<=\lambda_2<=\cdots<=\lambda_n$

如果 G 是 c 个连接部件，那么 L 有 c 个等于 0 的特征向量。如果 G 是连通的，$\lambda_2\neq 0$，λ_2 是 G 的连接代数值（Fiedler value）。其对应的特征向量为 Fiedler 向量。

当我们考虑 2-way 划分时，令 p 是 A 的划分指示向量：

$$p_j=\begin{cases}-1, j\in A^c\\ 1, j\in A\end{cases} \tag{8}$$

那么：$Cut(A,A^c)=f(p)=cut(A,A^c)=f(p)=\frac{1}{4}\sum_{i,j\in V}w_{ij}(p_i-p_j)^2=\frac{1}{2}p^TLp$

考虑约束 $x^TWe=x^TDe=0$，则

$$\min Ncut(A,A^c)=\min\frac{x^T(D-W)x}{x^TDx} \tag{9}$$

将 x 放松（松弛）到连续域$[-1,1]$，获得 minNCut 的问题就是：

$$\arg\min_x \min_{x^TDe=0}\frac{x^T(D-W)x}{x^TDx} \tag{10}$$

根据瑞利商原理，式(10)的优化问题等于下列等式的第二最小特征值的求解问题：

$$D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}x=\lambda x \tag{11}$$

对应于第二最小特征值对应的特征向量 x_2 则包含了图的划分信息。人们可以根据启发式规则在 x_2 寻找划分点 i，使得值大于等于 x_{2i} 的划为一类，而小于 x_{2i} 的划为一类。

同理，我们可以推理得到 k-way Ncut 目标函数式(6)的最优解在式(11)的 k 个最小特征值对应的特征向量所组成的子空间上。

2.3 算法描述

谱聚类算法由三个部分组成：

1. 建立表示样本集的矩阵 S。

2. 计算 S 的 k 个特征值与特征向量。

a. 2-way：将原始样本数据映射到一维空间（$k=1$）。

b. k-way：将原始样本数据映射到由 k 个正交向量的 k 维空间 S'。

3. 将 k 维子空间 S' 的行作为样本的新的数据表示，且基于这种新的表示，将样本进行聚类。

a. 2-way：在一维空间上根据目标函数最优原则划分。并且在划分好的两个子图上迭代划分。

b. k-way：利用传统的 k-means 或其它传统聚类算法在 k 维空间上进行聚类。

上述的描述是算法的一个框架，在具体的算法中，不同的算法在数据集矩阵 S 的表示上存在着不同，如：根据 2-way Ncut 的目标函数，$S=D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$；根据随机游动关系，$S=D^{-1}W$ 等。

3 当前的研究与应用

谱聚类方法最初是在图像分割中应用。Shi 和 Malik[2] 将像素作为顶点，根据像素的亮度和空间位置确定连接像素点间边的权值，利用 2-wayNcut 的谱聚类方法迭代地进行图的分割。该方法获得了满意的效果。

M. Gu, H. Zha 等人[15] 分析了在不规则图上进行 k-way 图划分的谱松散模型，并且根据该模型，提出了 k-way Ncut 与 k-way Min-Max cut 不同目标函数设置的下限值，同时分析了对应于最优解的特征空间或单个特征向量的代数结构，为将谱图理论运用到 k-way 图划分问题中，提供了理论基础。

Meila 和 Shi[4] 将相似性解释为 Markov 链中的随机游动，分析了这种随机游动的概率转移矩阵 $P=D^{-1}W$ 的特征向量（其中：W 是相似度矩阵），并且根据随机游动对 Ncut 进行了概率的解释，提出了基于随机游动的新的算法。同时，在这个解释框架下提出了多个特征相似矩阵组合下的谱聚类方法，在图像分割中也取得了很好的效果。

Zha 等人[5] 和 Dhillon[6] 研究了基于二分图 $G=\langle X,Y,W\rangle$ 上的谱聚类。聚类目标是使得在最小化二分图上的不匹配顶点对间的边权重和最小，故目标函数可以用变形的 Ncut 表示：

$$Ncut(V_1,\cdots,V_k)=\sum_{l=1}^{k}\frac{cut(X_1,Y_1^c)+cut(X_1^c,Y_1)}{\sum_{i\in X_1}\sum_j W_{ij}+\sum_i\sum_{j\in Y_1}W_{ij}} \tag{12}$$

将二分图的邻接矩阵 W 转换对称矩阵 $\widetilde{W}$：

$$\widetilde{W}=\begin{bmatrix}0 & W^T\\ W & 0\end{bmatrix}$$

然后再根据谱图理论对二分图上划分的目标函数式(12)进行分析（与 2 的分析相似）。Zha 等人与 Dhillon 发现最小化目标函数式(12)可以等同于与二分图相关联的边权重矩阵的奇异值分解。Dhillon[6] 将其运用到文档聚类中，对 CMU 的 Newsgroup20 做了实验，取得了很好的效果。基于二分图模型，该算法同样也可以用于市场分析中交易-商品的分析，生物信息挖掘中的 Gene expression profiles[7]。

Zha 等人[27] 分析了核 k-means 的方法，发现最小化核 k-means 的目标函数等同于一个由数据向量组成的 Gram 矩阵的迹最大化问题。同时，迹最大化问题的松散解可以通过 Gram 矩阵的部分特征分解获得。首次用谱松散的方法获得核 k-means 的目标函数的全局最优解。Dhillon[12] 在此基础上，又研究了加权核 k-means 的目标函数，将其与 Ncut 目标函数建立联系，提出了一个可以单调递减 Ncut 值的新颖的加权核 k-means 算法。

Ncut 是一个可行的聚类目标函数。它的求解是一个 NP 难问题。传统的方法是宽松的谱松散方法。Xing 与 Jordan[28] 则分析了对 Ncut 的半正定规划（SDP）模型。根据该模型，对 Ncut 提出了一个比谱松散更紧的下限。同时指出 Ncut 本身不能刻画最优的聚类，但它可以通过不同的松散方法获得合理的聚类。

谱聚类方法不仅用于无监督学习中，也用于有约束的半监督学习中。Kamvar 等人[26] 将 pageRank[30] 的随机游动模型运用到相似度矩阵中，根据已知样本的类别修正相似度矩阵。然后根据谱聚类算法获得聚类结果。Bach 与 Jordan[22] 则是根据一个基于已知的划分与 Ncut 谱松散结果的误差，提出了新的目标函数。通过最小化新的目标函数推出新的谱聚类算法，获得聚类结果。

4 关键问题和未来的研究方向

尽管谱聚类算法具有坚实的理论基础—谱图理论，并且在实践中也取得了很好的效果，但是它仍然存在一些关键问题：

1. 如何构造邻接矩阵 W：在谱聚类算法中，边的权值是顶点 i 与 j 的相似度 $sim(i,j)$，故表示图的邻接矩阵 W 也是样本空间的相似度矩阵。相似度矩阵的构造依赖于两个样本间相似度的构造。而单纯地依赖人为选择的相似度函数是带有一定的局限性的。我们应该引入领域知识，学习构造邻接矩阵。

2. 自动地确定聚类的数目：聚类数目的确定对聚类的质

量有很大的影响。当聚类数目大于实际聚类数，k-way 谱聚类方法的效果差[11]。因此如何自动地确定聚类数目是一个关键的问题，是未来研究的方向。

3. 如何解决模糊聚类的问题：尽管在文档聚类中，谱聚类取得了很好的效果。但是在文档聚类中，单个词可能属于多个类，单个文档可能是多主题的文档。这就需要我们用模糊聚类的方法解决。如何确定基于模糊聚类与谱方法的联合：如建立模糊标准的图划分的目标函数等，是我们的研究方向。

4. 运用到海量数据中去：当我们用谱聚类，不可避免地要计算矩阵的特征值与特征向量。通常这种计算的代价很大，求解非稀疏矩阵的所有特征向量的标准解法需要 $O(n^3)$。同时当应用到海量数据时，相似度矩阵也很大，可能会超出计算机的内存。Dhillon[12]在研究了谱聚类和核 kmeans 聚类方法的关系后，将谱聚类问题用核 k-means 算法求解，并且运用到核 k-means 算法的优化技巧[13,14]，来解决海量数据的计算。但该方法只用于非二分图的谱聚类方法，对二分图的聚类问题仍然是我们未来的研究方向。

参 考 文 献

1 Hagen L, Kahng A B. New spectral methods for ratio cut partitioning and clustering. IEEE Trans. Computer-Aided Design, 1992,11(9):1074～1085
2 Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000, 22(8):888～905
3 Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 14. MIT Press, 2001
4 Meila M, Shi J. A random walks view of spectral segmentation. In:8th International Workshop on Artificial Intelligence and Statistics, 2001
5 Zha H, He X, Ding C, Gu M, Simon H D. Bipartite Graph Partitioning and Data clustering. In: Proc. of ACM 10th Int'l Conf. Information and Knowledge Management (CIKM 2001), Atlanta, 2001. 25～31
6 Dhillon I S. Co-clustering documents and words using Bipartite Spectral Graph Partitioning. KDD 2001 San Francisco, California, USA
7 Ding H Q. Unsupervised feature selection via two-way orderign in gene expression analysis. http://bioinformatics. oupjournals. org/cgi/reprint/19/10/1259
8 Ding C, He X, Zha H, Gu Ming, Simon H. Spectral Min-Max Cut for Graph Partitioning and Data Clustering. In: Proc. of 1st IEEE Int'l Conf. Data Mining. San Jose, CA, 2001
9 Fieder M. Algebraic connectivity of graphs. Czechoslovak Math, J. 1973,22:298～305
10 Ding C, He X, Zha H. A spectral method to separate disconnected and nearly-disconnected Web graph components. KDD '01, San Francisco, California, USA
11 Verma D, Meila M. A Comarison of Spectral Clustering Algorithms:[UTCS Techniacl Report # TR-04-25]
12 Dhillon I, Guan Y, Kulis B. A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts:[UTCS Technical Report # TR-04-25]
13 Zhang R, Rudnicky A. A large scale clustering scheme for kernel k-means. In ICPR02, 2002. 289～292
14 Dhillon I S, Guan Y, Kogan J. Iterative clustering of high dimensional text data augmented by local search. In: Proceedings of The 2002 IEEE International Conference on Data Mining, 2002
15 Cu M, Zha H, Ding C, He X, Simon H, Xia J. Spectral relaxation models and structure analysis for K-way graph clustering and bi-clustering. http://citeseer. st. psu. edu/context/2380189/639394
16 Weiss Y. Segentation using eigenvectors: a unifying view. http://www. cs. huji. ac. il/～yweiss/iccv99. pdf
17 Meila, Xu L. Multiway Cuts and SPectral Clustering. http://www. stat. washington. edu/www/research/reorts/
18 Belkin M, Niyogi P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, Advances in Neural Information Processing Systems 14 (NIPS 2001)
19 Yu S U, Shi J. Multiclass spectral clustering. In: Int'l Conf. on Comuter Vision, 2003
20 Kannan R, Vempala S, Vetta A. On clusterings - good, bad, and spectral. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, 2000
21 Chan P K, Schlag M, Zien J Y. Spectral k-way ration-cut partitioning and clustering. IEEE Tarns. CAD-Integrated Circuits and Systems, 13:1088～1096
22 Bach F R, Jordan M I. Learning spectral clustering. Neural Info. Processing Systems 16 (NIPS 2003), MIT Press, 2004
23 Brand M, Huang K. A unifying theorem for spectral embedding and clustering. In: Int'l Workshop on AI & Stat (AI-STAT 2003), 2003
24 Ding C, He X, Zha H, Simon H. Unsupervised learning: self-aggregation in scaled principal component space. In: Proc. 6th European Conf. Principles of Data Mining and Knowledge Discovery (PDKK 2002), 2002. 112～124
25 Ding C. Dovument Retrieval and Clustering: from Principal Compongent Analysis to Self-aggregation Networks. In: Int'l Workshol on AI & Stat(AI-STAT 2003), 2003
26 Kamvar S D, Klein D, Manning C D. Spectral Learning. IJCAI-03, 2003
27 Chung F R K. Spectral Graph Theory. Amer. Math. Soc, 1997
28 Xing E P, Jordan M I. On semidefine relaxation for normalized k-cut and connections to spectral clustering. http://www. cs. berkeley. edu/～epxing/Paper/TR-SDP5. ps
29 Zha H, He X, Ding C, Cu G, Simon H. Spectral relaxation for K-means clustering. http://www. cs. psu. edu/～ zha/papers/ming. ps
30 Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proc. 7th International World Wide Web Conference, 1998

(上接第 197 页)

量中除了最大值的权重外其它值权重都为 $1/n^2$ 时，则最小加权平均算子变为等权平均算子。

性质 3(归一性) n 维最大、最小加权平均算子的新权系数都是 n 维权重向量。

定理 1 最大最小加权平均算子与加权平均算子 WM，最小值算子 Min，最大值算子 Max 有如下关系：

$$Min \leqslant W_{\omega}^{\min} \leqslant WM \leqslant W_{\omega}^{\max} \leqslant Max \qquad (7)$$

定理 1 说明算子 $W_{\omega}^{\max}$ 和 $W_{\omega}^{\min}$ 是介于 Max 算子，加权平均算子 WM 和 Min 算子之间的算子。算子 $W_{\omega}^{\max}$ 和 $W_{\omega}^{\min}$ 将最大最小值算子与加权平均算子进行了泛化，综合了最大最小值算子与加权平均算子的优点。

结束语 本文提出的加权模糊聚集算子适合于所有的模糊集合与模糊信息源的聚集运算，在数据挖掘、模糊专家系统、模糊控制系统、多 Agent 系统以及数据集成领域等均有着广泛应用前景。我们将其应用到了开发的专家系统系统平台中，对复合模糊命题的计算与多结果的聚集运算在实践过程中得到了良好的应用。进一步工作是根据本文的理论研究成果，针对不同的需求研究合适的具体变权算子形式。

参 考 文 献

1 Yager R R. On weighted median aggregation. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 1994. 101～113
2 Domingo J, Torra V. Median-Based Aggregation Operators for Prototype Construction in Ordinal Scales. International Journal of Intelligent System, 2003, 18: 633～655
3 Han J W, Kamber M. Data Mining: Concepts and Techniques. San Mateo, CA: Morgan Kaufmann, 2000
4 Li A P, Wu Q Y. On Harmonic Triangular Norm Aggregation Operators in Multicriteria Decision. In: Proceedings of the 8th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, 2004. 66～71
5 Yager R R. Induced aggregation operators. Fuzzy Sets Systems, 2003
6 Yager R R. On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans Syst Man Cybernet, 1988, 18:183～190
7 何新贵. 模糊知识处理的理论与技术(第 2 版). 北京：国防工业出版社，1999
8 陆志峰. 模糊逻辑的研究，计算机工程与应用，1999,8:15～19
9 汪培庄，李洪兴，模糊系统理论与模糊计算机. 北京：科学出版社，1996
10 Dubois D, Prade H. Weighted minimum and maximum operations in fuzzy sets theory. Information Sciences, 1986, 39:205～210