

文章编号:1001-9081(2012)12-3274-04

doi:10.3724/SP.J.1087.2012.03274

面向大文本数据集的间接谱聚类

侯海霞^{1*}, 原民民², 刘春霞³

(1. 太原大学 计算机工程系, 太原 030032; 2. 山西水利职业技术学院 信息工程系, 山西 运城 044000;

3. 太原科技大学 计算机科学与技术学院, 太原 030024)

(* 通信作者电子邮箱 houhaixia@163.com)

摘要:针对谱聚类存在计算瓶颈的问题,提出了一种快速的集成算法,称为间接谱聚类。它首先运用 K-Means 算法对数据集进行过分簇,然后把每个过分簇看成一个基本对象,最后在过分簇的级别上利用标准谱聚类来完成总体的聚类。将该思想应用于大文本数据集的聚类问题后,过分簇中心之间的相似性度量方法可以采用常用的余弦距离法。在 20-Newgroups 文本数据上的实验结果表明:间接谱聚类算法在聚类准确性上比 K-Means 算法平均高出 14.72%;比规范割谱聚类仅低 0.88%,但算法所需的计算时间平均不到规范割谱聚类的 1/16,且随着数据集的增大当规范割谱聚类遭遇计算瓶颈时,提出的算法却能快速地给出次优解。

关键词:谱聚类;文本聚类;大数据集

中图分类号: TP301.6 **文献标志码:**A

Indirect spectral clustering towards large text datasets

HOU Hai-xia^{1*}, YUAN Min-min², LIU Chun-xia³

(1. Department of Computer Engineering, Taiyuan University, Taiyuan Shanxi 030032, China;

2. Department of Information Engineering, Shanxi Conservancy Technical College, Yuncheng Shanxi 044000, China;

3. College of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan Shanxi 030024, China)

Abstract: To alleviate the computational bottleneck of spectral clustering, in this paper a general ensemble algorithm, called indirect spectral clustering, was developed. The algorithm first grouped a given large dataset into many over-clusters and then regarded each obtained over-cluster as a basic object. And then the standard spectral clustering ran at this object level. By convention, when applying this new idea to large text datasets, the cosine distance would be the appropriate manner in measuring the similarities between over-clusters. The empirical studies on 20-Newgroups dataset show that the proposed algorithm has a 14.72% higher accuracy on average than the K-Means algorithm and has a 0.88% lower accuracy than the normalized-cut spectral clustering. However, the proposed algorithm saves 16.8 times computation time compared to the normalized-cut spectral clustering. In conclusion, with the increase of data size, the computation time of the normalized-cut spectral clustering might become unacceptable; however, the proposed algorithm might efficiently give the near-optimal solutions.

Key words: spectral clustering; text clustering; large dataset

0 引言

随着 Web 2.0 的发展,网络的交互性日益加强,每天都有大量的博客、新闻和评论等文档数据产生,这些网络文本数据的快速积聚给传统的数据挖掘算法带来巨大的挑战,迫切需要提出快速而有效的数据挖掘算法发现数据中存在的知识。另一方面,监督信息的获取往往耗费大量的人力和时间,因此文本挖掘的主要任务往往是无监督分类,即聚类。

聚类的目的是探测数据的内部结构,将相似的数据划分到同一组,将不相似的数据划分到不同的组。聚类技术已经经历了半个多世纪的发展,研究学者们提出了形形色色的聚类算法,包括:基于中心的聚类、基于密度的聚类、基于图论的聚类、基于网格的聚类和基于模型的聚类。然而,每一类聚类算法都不是通用的,它们分别适合于不同的应用场景,各有优缺点,例如:基于中心的算法 K-Means^[1]计算代价较低,但是应用范围较窄,适合于处理球状的簇结构并经常收敛到局部

解;基于密度的算法^[2]可以处理任意形状的簇结构,但是参数往往难以设置;基于网格的聚类算法等^[3]的优点是处理速度很快,但具有较低的聚类准确性;基于模型的聚类高斯混合算法^[4]可以用于处理不完全数据,然而它仅仅是假设数据是高斯分布的。

近年来,最受研究者关注的一类聚类算法是谱聚类^[5-14]。它属于基于图论的聚类算法,更确切地讲,它是基于谱图理论^[15]的一类算法。谱聚类的基本思想是将聚类问题看成是图的划分问题,并利用谱放松方法来实施图的分割。此类算法不关心数据的初始分布,对于数据簇结构的形状具有较好的适应性。实验结果表明,在大多数基准数据集上,谱聚类的聚类质量比最常用的 K-Means 算法要好。K-Means 算法或者高斯混合算法的计算速度较快,因为它们只需要比较数据点离中心的距离。然而,这些算法隐含着重大的缺陷,即假设每个簇具有球状或者高斯分布。谱聚类运用的是成对距离,而不是数据点到中心点的距离,它突破了中心式聚类的限

收稿日期:2012-07-23;修回日期:2012-09-03。 基金项目:山西省青年科技研究基金资助项目(2011021014-3)。

作者简介:侯海霞(1978-),女,山西阳泉人,讲师,硕士,主要研究方向:软件工程、算法分析; 原民民(1977-),男,山西运城人,讲师,硕士,主要研究方向:算法分析、计算机程序设计; 刘春霞(1977-),女,山西大同人,讲师,硕士,主要研究方向:计算机智能控制及系统优化。

制,更加注重数据所反映的流形结构。谱聚类也有明显的问题,成对距离的计算和存储往往耗费 $O(n^2)$ 的复杂性,对于中、大规模的聚类问题,存储空间耗费巨大。

若能将基于图划分的谱聚类算法扩展到大数据集,则可以继承谱聚类聚类准确性高的优势,因此这种扩展具有重要的现实意义。在这个问题上,研究学者们已经取得了一些成果。例如:对于图像分割的情形,Fowlkes 等^[16]提出了运用 Nyström 方法实施特征向量的低秩近似;Dhillon 等^[17]避免求解特征值问题,并采用多级聚类算法实现图的划分;Song 等^[18]提出了并行谱聚类算法,可以处理 10 万数量级以上的问题。本文算法与上述方法均不同,Fowlkes 算法是利用一个随机子集作为处理对象,而本文算法是利用代表点集作为处理对象的,代表点的选择运用的是 K-Means 算法。Dhillon 的多级方法避免了特征分解,实际上,采用了 Lanczos 迭代法^[19]后,特征分解并不是谱聚类问题的主要瓶颈,真正的瓶颈是相似度矩阵的构造时间和存储代价,本文算法将继承谱宽松方法的基本思想。另外,本文算法仅仅运行在单 PC 机上。

1 谱聚类的基本知识

谱聚类是一类算法,拥有几种不同的版本。本文仅以最常用的规范割谱聚类为例进行介绍。

1.1 基本概念

给定一个向量数据集 $X = \{\mathbf{x}_i\}_{i=1}^n$,首先需要选定一个相似性度量标准,用于度量成对向量数据之间的相似度,记作: $Aff(\mathbf{x}_i, \mathbf{x}_j)$,度量标准的选择取决于数据集类型。本文关心的是文本聚类,常用向量之间的夹角的余弦来表示其相似度。在此基础上,需要构造一个 $n \times n$ 的相似度矩阵来反映数据之间的两两相似关系,将该矩阵记为 A ,该矩阵对应的图称为相似图。当然,为了减小存储代价,构造的矩阵也可以具有某种稀疏性,例如采用 k 近邻图。

1.2 规范割谱聚类

既然谱聚类是一种基于图划分的聚类算法,就需要定义某种划分标准。早期的最小割算法难以保证聚类的均衡性,因此有了后来的比率割、规范割、最小最大割等划分标准。下面以规范割(Normalized cut, Ncut)^[5]为例来加以介绍:令 V 表示相似图的顶点集, V_j 表示第 j 个顶点子集, $W(\cdot, \cdot)$ 表示两个顶点集之间的边权总和,即相似度的总和,则规范割 $Ncut$ 可写为如下表达式:

$$Ncut = \sum_{j=1}^m \frac{W(V_j, V) - W(V_j, V_j)}{W(V_j, V)} = \sum_{j=1}^m \frac{cut(V_j, V \setminus V_j)}{vol(V_j)} \quad (1)$$

若期望的聚类数为 K ,则规范割谱聚类等价于:将向量数据集 X 划分为 K 类,使得 $Ncut$ 的值最小化。进一步,该问题也可以写成如下形式:

$$\min_{U \in \mathbb{B}^{n \times k}} \text{trace}(U^T L U) \quad (2)$$

$$\text{s. t. } U^T U = I$$

其中: L 表示的规范拉普拉斯矩阵; $\text{trace}(\cdot)$ 表示求矩阵的迹; U 是要求解的二值指示矩阵,其每行仅有一个元素为 1,其余为 0,同一列中的元素值为 1 的行编码的数据点表示划分到同一个类中。然而,该问题是典型的离散最优化问题,属于 NP 问题。为了快速求解的需要,问题^[8]需要先放松为:

$$\min_{U \in \mathbb{R}^{n \times k}} \text{trace}(U^T L U) \quad (3)$$

$$\text{s. t. } U^T U = I$$

即允许 U 在实数域取值。而上述问题等价于求解规范拉氏矩阵 L 的最小的 K 个特征值对应的特征向量,这些特征向量组成了实数域的 U ,最后将 U 的每行看成一个数据点,对所有新数据点利用 K-Means 算法进行聚类。由于每个新数据点与一个原始空间的数据点相对应,故而新数据点的标签即是原始数据点的标签。

2 间接谱聚类算法及理论分析

如前所述,将谱聚类“直接”用于大数据集是不明智的,它会导致存储和计算的瓶颈。因此,需要“间接”使用谱聚类算法:首先对数据进行预处理,抽取代表点,然后在代表点级别上进行谱聚类。具体来讲,首先运用 K-Means 算法将原始数据集划分成许多小簇,这些小簇称为过分簇,使得过分簇的数量远大于真实簇数,然后将每个过分簇用其簇中心作为代表点,在代表点的级别上运行谱聚类算法。本文算法框架描述如下:

- 输入 向量数据集 X ,过分簇数 c ,期望的簇数 K ;
- 输出 簇指示向量。
- 步骤 1 首先在向量数据集 X 上运行 K-Means 算法,将向量数据集聚为 c 个过分簇;
- 步骤 2 将每个过分簇的中心作为一个代表点,构造一个 $c \times c$ 的相似度矩阵 R ;
- 步骤 3 在 R 上运行规范割谱聚类;
- 步骤 4 令过分簇中的向量数据拥有与过分簇中心一致的标签。

仔细分析上述算法,不难发现,算法隐含着一种假设,即:过分簇是球状的。理论上来讲,这个假设是容易成立的。下面通过举例来阐释该观点。图 1 所示的数据集是一个包含两个流形簇的 2 维数据集,若直接采用 K-Means 算法进行聚类,则得到的聚类结果是:虚线左上侧的点被看成一类,虚线右下侧的点被看成另一类。这种划分明显不是人们所期望的,因为有 8 个数据点被错误地划分。当然,直接运用 K-Means 算法之所以得到不满意的结果,原因是由于算法假设簇形状是球状的,而实际的簇形状非球状。而本文算法首先要采用 K-Means 算法作为预处理的手段。图 2 所示是过分聚类的结果:利用 K-Means 将原始数据集划分成 12 个过分簇,很容易看出,每个过分簇中的点具有相似的性质,或者说包含在同一个流形簇中,没有错分的情况出现。在这个例子中,过分聚类的预处理操作是有效的。

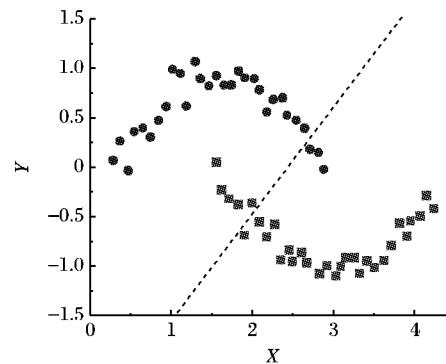


图 1 直接运用 K-Means 聚类的结果

接下来分析参数 c 的取值,这一点主要从算法的计算代价上来考虑。首先,K-Means 过分聚类的时间复杂性为

$O(cm^2)$, 其中, m 表示 K-Means 的迭代次数, n 表示向量数据的个数, c 表示过分簇的个数。为了保证算法的计算速度快, c 的取值应该远小于 n 。同时, 为了使得过分簇内的数据点尽量具有相同的性质, c 的取值应该远大于真实簇数 K 。例如: 对于一个包含 10 类的 2 万个文档的数据集, c 取几百应该是合理的。而且当 c 在 10^2 数量级时, 后续的聚类算法只需要将一个小的相似度矩阵调入内存, 并进行后续的特征方程求解, 这些均可以快速计算出结果。划分式聚类通常假设真实聚类数 K 是可估计的或者已知的, 在实验中, 由于所采用的数据为基准公开数据, 其类别是已知的, 故而 K 将设置为真实类别的个数。

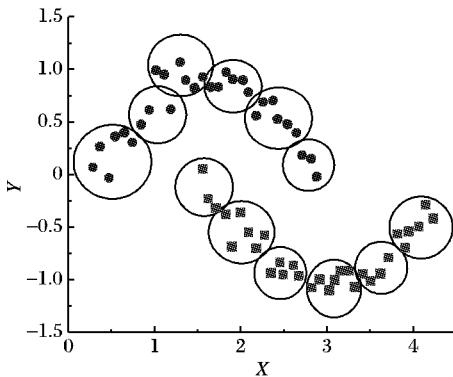


图 2 运用 K-Means 进行过分聚类的结果

对于每个给定的文档, 需要运用向量空间模型将其转化为一个实向量, 文档之间的关系于是转化为实向量之间的关系。通常, 余弦距离比较适合度量文本之间的相似度, 因此, 在步骤 2 中, 相似度矩阵的构造采用余弦距离的度量方式。当然, 本文的重点是算法的改进, 而不关心文档是如何转化成实向量的。

3 实验评估

下面利用实验来评估本文算法。目的是回答这样三个问题:

- 1) K-Means 过分聚类是否能够使得相似的数据尽量落入同一个过分簇中。
- 2) 相对于 K-Means 算法, 本文算法是否在聚类准确率上有一定的优势。
- 3) 本文算法是否在聚类准确率上与规范割谱聚类接近, 并能节省较多的运算时间。

后续的实验紧紧围绕这三个问题来展开。

3.1 基准数据集

本文采用的基准数据集是 UCI 机器学习库^[20]的文档数据(20-Newsgroups), 从中选择 4 大类 17 小类文档子集作为实验数据, 它们的类别信息如下:

```
comp. graphics
comp. os. ms-windows. misc
comp. sys. ibm. pc. hardware
comp. sys. mac. hardware
comp. windows. x
rec. autos
rec. motorcycles
rec. sport. baseball
rec. sport. hockey
sci. crypt
```

```
sci. electronics
sci. med
sci. space
talk. politics. guns
talk. politics. mideast
talk. politics. misc
talk. religion. misc
```

在实验中仅将其看成 4 个大类, 其标签分别是: comp.* , rec.* , sci.* , talk.* , 一共包含 16 242 个文档, 该数据子集可以从 <http://cs.nyu.edu/~roweis/data.html> 下载, 文档集已经利用向量空间模型转化为向量集, 向量的维度是 100。

3.2 聚类评价准则

本文选择的 UCI 数据集的标签是已知的, 因此可以基于真实标签来评价聚类算法的结果。对于过分聚类算法和正常的聚类算法, 其评价标准需要分别考虑。

对于过分聚类算法, 需要定义一个新的评价准则。令过分聚类数为 c , 真实类别数为 K , 则可以构造一个 $K \times c$ 的矩阵 G , 当执行完过分聚类算法之后, 对 G 进行循环赋值, 使得 G_{ij} 表示真实类标签为 i 的文档被指派到第 j 个过分簇中的文档数。进一步, 过分聚类的准确率(Accuracy 1, 缩写为 ACC1)评价方法可以定义如下:

$$ACC1 = \sum_{j=1}^c \max\{G_{i,j}\} / n \quad (4)$$

其中 $G_{i,j}$ 表示的是矩阵 G 的第 j 列元素。

对于一般的聚类算法, 其对应的矩阵 G 是一个 $K \times K$ 的方阵。求解其准确率的问题等价于: 在矩阵 G 的每行每列只取一个元素, 并将这些元素相加, 使得和最大化, 这里将最大和记作 s 。这个问题等价于二分图的最大匹配问题, 可以利用匈牙利算法来求解。于是聚类的准确率(Accuracy 2, 缩写为 ACC2)可以表达为:

$$ACC2 = s/n \quad (5)$$

3.3 实验结果

K-Means 过分聚类算法是本文算法的一个重要的数据预处理步骤, 为了检查该步骤的有效性, 首先通过一组文档数据来观察过分聚类数的增加对过分聚类质量的影响。选用 sci.* 和 talk.* 共 8 118 个文档, 分别设置参数 $c = 2, 20, 40, 80, 160, 200, 240$ 。图 3 展示了 ACC1 随 c 的变化趋势。容易看出, 随着参数 c 的增大, 过分聚类的准确性呈总体变高的趋势。这意味着过分簇的同质性呈增长的态势, 过分簇的高同质性是本文算法有效性的基础。

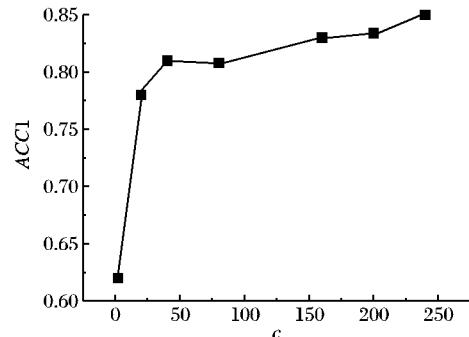


图 3 K-Means 过分聚类的准确率随参数 c 的变化趋势

再来检查本文算法的效果。实验中, 将 20-Newsgroup 数据的 4 个大类进行了 5 种不同的组合, 具体的方案见表 1。实验的硬件配置为普通 PC 机(CPU 为 Intel T2130 1.86 GHz, 内存 2 GB)。对于所有数据组合, 参数 c 均设置为 160。分别

运行 K-Means 算法、本文算法和规范割谱聚类算法,表 1 列出了各自的聚类准确率 ACC2 和运行时间。可以看出,本文算法的聚类准确性大大超过 K-Means 算法。特别是在第 1 组数据集上,准确率的提升幅度达到 19.4%。本文算法的计算开销虽然比 K-Means 算法大,但它是可以接受的,这些计算开销主要消耗在了过分聚类步骤上。此外,为了比较的需要,表 1 也

列出了正常的规范割谱聚类的聚类准确率和计算时间,可以看出,虽然其聚类准确性轻微地优于本文算法,但是在时间的开销上远大于本文算法。而且,理论上,随着数据的继续增长,直接的规范割谱聚类算法将变得不再适用,而本文的间接谱聚类算法却仍然可以正常使用,这说明本文算法具有更强的伸缩性。

表 1 聚类准确性和运行时间比较

数据组合	文档数	类别数	K-Means		间接谱聚类		规范割谱聚类	
			ACC2	时间/s	ACC2	时间/s	ACC2	时间/s
sci, talk	8 118	2	0.614	0.752	0.808	55.8	0.821	597.2
comp, rec	8 124	2	0.608	0.609	0.807	43.8	0.811	673.7
comp, rec, sci	10 781	3	0.439	1.080	0.532	68.6	0.537	1 091.6
rec, sci, talk	11 637	3	0.431	1.330	0.591	93.3	0.592	1 239.4
rec, sci, talk, comp	16 242	4	0.544	2.080	0.634	100.6	0.655	2 482.8
平均			0.5272	1.1702	0.6744	72.42	0.6832	1 216.94

4 结语

谱聚类具有处理复杂聚类结构的能力,但是其计算存在着瓶颈,难以直接应用到大数据集上。针对这一问题,提出了先用 K-Means 算法抽代表点,后执行谱聚类的集成算法。理论分析的结果是,这一思路可能带来比 K-Means 更好的聚类准确率,其计算开销较直接的谱聚类大大减少。在文本基准数据上的实验结果也验证了本文算法的有效性。

参考文献:

- [1] MACQUEEN J B. Some methods of classification and analysis of multivariate observations [C] // Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281 – 297.
- [2] ESTER M, KRIEGEL H, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C] // Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Oregon: AAAI Press, 1996: 226 – 231.
- [3] WANG W, YANG J, MUNTZ R R. STING: A statistical information grid approach to spatial data mining [C] // Proceedings of the International Conference on Very Large Data Bases. Athens: AAAI/MIT Press, 1997: 186 – 195.
- [4] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society: Series B, 1977, 39(1): 1 – 38.
- [5] SHI J B, MALIK J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888 – 905.
- [6] ALZATE C, SUYKENS J A K. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(2): 335 – 347.
- [7] MAIER M, LUXBURG U, HEIN M. Influence of graph construction on graph-based clustering measures [C] // Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2009: 1025 – 1032.
- [8] LUXBURG U. A tutorial on spectral clustering [J]. Statistics and Computing, 2007, 17(4): 395 – 416.
- [9] CHEN W, SONG Y, BAI H, et al. Parallel spectral clustering in distributed systems [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 568 – 586.
- [10] RANGAPURAM S S, HEIN M. Constrained 1-spectral clustering [J]. Journal of Machine Learning Research, 2012, 22: 1143 – 1151.
- [11] YAN D, HUANG L, JORDAN M I. Fast approximate spectral clustering [C] // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 907 – 916.
- [12] LI M, LIAN X C, KWOK J T, et al. Time and space efficient spectral clustering via column sampling [C] // Proceedings of International Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2011: 2297 – 2304.
- [13] FILIPPONE M, CAMASTRA F, MASULLI F, et al. A survey of kernel and spectral methods for clustering [J]. Pattern Recognition, 2008, 41(1): 176 – 190.
- [14] CHEN XINLEI, CAI DENG. Large scale spectral clustering with landmark-based representation [EB/OL]. [2012-05-20]. <http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/download/3484/3883>.
- [15] CHUNG F. Spectral graph theory [EB/OL]. [2012-05-20]. <http://www.math.ucsd.edu/~fan/research/revised.html>.
- [16] FOWLkes C, BELONGIE S, CHUNG F R K, et al. Spectral grouping using the nystrom method [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(2): 214 – 225.
- [17] DHILLON I S, GUAN Y, KULIS B. Weighted graph cuts without eigenvectors: a multi-level approach [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(11): 1944 – 1957.
- [18] SONG Y, CHEN W, BAI H, et al. Parallel spectral clustering [C] // Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Part II. Berlin: Springer-Verlag, 2008: 374 – 389.
- [19] CULLUM J K, WILLOUGHBY R A. Lanczos algorithms for large symmetric eigenvalue computations: documentation and listings original Lanczos codes [EB/OL]. [2012-05-20]. http://www.netlib.org/lanczos/vol2/Chp_1_Overview.pdf.
- [20] FRANK A, ASUNCION A. UCI machine learning repository [EB/OL]. [2012-05-20]. <http://archive.ics.uci.edu/ml>.