全国大学生数学建模竞赛专栏

碎纸片拼接复原的数学方法

薛毅

(北京工业大学 应用数理学院, 北京 100124)

摘　要:就 2013 年"高教社杯"全国大学生数学建模竞赛 B 题"碎纸片的拼接复原"提出了一种用"纯数学手段"完成拼接复原的方法,可概括为 3 步:TSP,聚类分析和双面信息的利用。根据题目要求给出了 3 个步骤中人工干预的方式与时间节点。

关键词:旅行商问题;聚类分析;碎纸片拼接

中图分类号:O221.7;O212.4;TP391.41　　**文献标志码**:A　　**文章编号**:2095-3070(2013)5-6-0009-05

碎纸复原问题最著名的例子应该是 2011-10-29 美国国防部高级研究计划局宣布的碎纸复原挑战赛。该项赛事共有 9000 多支队伍参加,获胜队将得到 50 000 美元的奖金[1]。今年的竞赛题"碎纸片的拼接复原"看似与挑战赛相同,但问题的难度有着本质的不同。首先,需要拼接复原的文件是纯文本的打印文件,并不是挑战赛的手写文件;其次,碎纸片由计算机生成,也就是说,碎纸片没有毛边,并不是真正由碎纸机粉碎而成的,这样就大大降低了问题的难度,使学生有可能在 3 天的时间内完成。为了确保学生在 3 天的时间内完成竞赛内容,命题人特意将问题分解成 3 个子问题:

问题 1　仅考虑只有纵切情形的碎纸片;

问题 2　碎纸片是由纵切与横切两种情形得到的;

问题 3　双面打印文件碎纸片的拼接复原。

这 3 个问题实际上在告诉我们解题的思路,因此,解决问题的方案也分成 3 步:1)建立仅有纵切情形的数学模型—旅行商问题(traveling salesman problem, TSP);2)先将碎纸片还原到各自所在的行—聚类分析,再使用 1)中的方法(TSP)复原成正确的"横条",最后使用人工干预的方法拼接成原始文件;3)使用碎纸片的双面信息,提高行聚类的准确率。根据题目要求,在这 3 步中均要考虑人工干预的方式,以及干预的时间节点。

1　仅有纵切的情形—旅行商问题

1.1　数据的读取与处理

读取数据是求解问题的关键。读取数据的方法是非常简单的,大部分的计算机语言都可完成,例如 Matlab 软件中的 imread 函数。

读取数据后,可以将数据作二值化处理。直观来讲,二值化处理的目的是使黑的更黑,白的更白,便于后面的聚类与拼接。笔者曾做过经二值化和未经二值化处理的数据的对比实验,结果表明,数据经二值化处理后其复原率有一定程度的提高。

1.2　建模—旅行商问题

建模思想是基于"相邻的两条碎纸片的灰度应该比较接近"这一事实提出的,它同时也给出了完成问题 1 求解的启发式算法。

1) 找出整个文件中最左侧的碎纸片(碎片的左侧为空白,数值均为 255,设碎纸片编号为 k_1,置 index =

收稿日期:2013-11-13

通讯作者:薛毅,E-mail:xueyi@bjut.edu.cn

k_1，$I=\{1,2,\cdots,19\}\backslash\{k_1\}$，$i=1$；

2）如果 $i=19$，则停止计算，输出拼接复原图序号 index，否则计算第 $j(j\in I)$ 个碎纸片最左侧的列与第 k_i 个碎纸片最右侧的列之间的距离，记距离最小的碎纸片的编号为 k_{i+1}；

3）置 index = index $\cup$ k_{i+1}，$I=I\{k_{i+1}\}$，$i=i+1$，转 2）。

由于仅纵切情形的碎纸片两侧的信息量较大，上述启发式算法可以得到附件1和附件2的拼接复原图，而且也不需要人工干预。

这种方法看似很好，但也存在两个致命的缺点：第一，它是一种局部寻优方法，而且计算复杂度高；第二，这种方法不便于推广，不利于在问题2和问题3的求解中使用。

为克服上述算法的缺点，提出一种全局寻优的方法，即将问题转化成旅行商问题（TSP）。将碎纸片看作城市（共19个），定义城市间的距离，即碎纸片之间的距离，定义方法为：对于碎纸片A和碎纸片B，用碎纸片A最右侧的列与碎纸片B最左侧的列之间的距离定义为两者之间的距离，同理可以定义碎纸片B到碎纸片A的距离。用此方法可以定义两个碎纸片之间的距离，形成一个一个非对称的距离矩阵。注意到，整个文件最左侧碎纸片的左侧与最右侧碎纸片的右侧均为空白，这样就可以形成圈。因此，仅纵切情形的拼接复原问题转化求解最小距离的 Hamilton 圈，即旅行商问题。

1.3　距离的定义

距离的定义有多种，如果数据没有二值化，可以使用绝对值距离（或称棋盘距离）、Euclid 距离、Chebyshev 距离等。二值化后的数据可以使用 Hamming 距离或 Jaccard 距离。本文使用的是 Jaccard 距离，定义为 $d_{ij}=\dfrac{m_2}{m_1+m_2}$，其中，$m_1$ 为两个向量中 1-1 配对的总数，m_0 为 0-0 配对的总数，m_2 为不配对的总数，$m_0+m_1+m_2$ 为向量的维数。

1.4　TSP 求解

旅行商问题是NP难度的，求解可能会遇到困难。但本问题的规模较小，只有19条碎纸片，所以得到TSP的最优解是相当顺利的。关于 TSP 的求解算法有很多，求解时可以选择任何一种（因为数据量不大），但需要注意的是，要选择那些能够处理求解非对称距离矩阵的方法或程序包，例如 LINGO 软件[2]。

1.5　图形复原

注意到最左边碎纸条的左侧和最右边碎纸条的右侧为空白，将得到的 Hamilton 圈从此处剪开，得到复原图。由于空白处的数值是255，剪切工作可由计算机自动完成，所以问题1的拼接复原工作并不需要人工干预，而且附件1（中文）和附件2（英文）的复原率均可达到100%。

2　带有横切的情形 — 聚类分析方法

问题2是讨论既有纵切又有横切的碎纸片的拼接复原。如果每个碎纸片能找到它应在的行，这样就将问题简化成仅有纵切的情况，可以使用问题1的模型与方法进行拼接复原。

下面讨论如何使每个碎纸片找到它应在的行，这实际上是一个聚类问题。注意到文本文件是黑白相间的，原属同一行的碎纸片文件黑白之间的间隔是可以对齐的。利用这一特点建立聚类的数学模型，将所有碎纸片聚成11类。如果正确完成这一步工作，就可以通过求解旅行商问题得到11个横条，剩下的工作由手工拼接就可以完成。

2.1　聚类分析中距离的定义

在读取数据后，每一碎纸片的数据是一个矩阵，对矩阵的每一行求均值（或求和，是等价的），将矩阵转换成向量，再定义两个向量之间的距离。除前面提到的各种距离外，还可以使用相关系数定义距离，在碎纸片的聚类分析中，这种距离是较为有效的。

2.2　聚类方法与人工干预

对于聚类方法，可以采用教科书介绍的系统聚类法，如最短距离法，最长距离法，中间距离法，类平均法，重心法和离差平方和法（Ward 方法）。经过对上述聚类方法做数值实验对比得知，效果最好的是离差平方和法，因为它有如下性质：如果分类正确，则同类样本之间的离差平方和较小，不同类样本之间的离差平方和较

大，而且每个类中样本的个数比较均匀。具体计算可以选择 Matlab 统计工具箱中的聚类分析函数[3]：pdist 函数，linkage 函数和 cluster 函数。

注意到原属同一行的碎纸片文件黑白之间的间隔是相互对齐的，但反过来，黑白之间的间隔相互对齐的碎纸片可能并不属于同一行。所以聚类后需要根据计算结果作适当的人工干预，将不属于某行的碎片文件去掉，有可能的话，放到它应在的行中。

3.2.1　中文文件的处理

对于中文文件（附件 3 中的碎片），由于各行的黑白间隔比较明显，只要采用前面介绍的常规处理方法就可以得到较为满意的聚类结果，表 1 给出的是附件 3 中碎纸片的聚类结果。

表 1　中文碎片的聚类结果

类号	碎片编号																			
1	1	18	23	26	30	41	50	62	76	86	87	100	120	142	147	168	179	191	195	
2	3	12	14	31	39	51	73	82	107	115	128	134	135	159	160	169	176	199	203	
3	34	42	43	47	58	77	84	90	94	97	112	121	124	127	136	144	149	164	183	
4	8	9	24	25	35	38	46	74	81	88	103	105	122	130	148	161	167	189	193	
5	2	11	22	28	49	54	57	65	91	95	118	129	141	143	178	186	188	190	192	
6	5	10	29	37	44	48	55	59	64	75	92	98	104	111	171	172	180	201	206	
7	6	19	20	36	52	61	63	67	69	72	78	79	96	99	116	131	162	163	177	
8	0	7	32	45	53	56	68	70	93	126	137	138	153	158	166	174	175	196	208	
9	13	15	17	27	33	60	71	80	83	85	125	132	133	152	156	165	170	198	200	202	205
10	4	40	89	101	102	108	113	114	117	119	123	140	146	151	154	155	185	194	207	
11	16	21	66	106	109	110	139	145	150	157	173	181	182	184	187	197	204			

在聚类计算中，使用相关系数定义各碎纸片之间的距离，采用离差平方和法作为系统聚类方法。从表 1 中可以看出，只有 9 号类和 11 号类出现差错，9 号类多了两个碎纸片，而 11 号类少了两个。直接对 9 号类碎纸片使用 TSP 方法拼接复原，其结果如图 1 所示。从图形容易看出，粗线以外的两个碎纸片应该不在该行上。

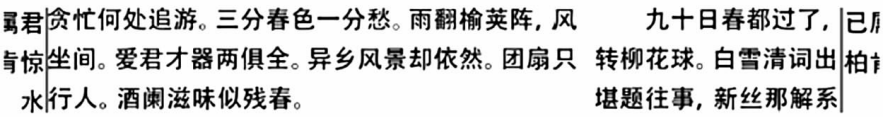

图 1　9 号类复原图

将这两个碎纸片从 9 号类中去掉，并放在 11 号类中，然后使用 TSP 方法拼接复原，结果如图 2 和图 3 所示。

九十日春都过了，贪忙何处追游。三分春色一分愁。雨翻榆荚阵，风转柳花球。白雪清词出坐间。爱君才器两俱全。异乡风景却依然。团扇只堪题往事，新丝那解系行人。酒阑滋味似残春。

图 2　人工干预后的 9 号类复原图

已属君家。且更从容等待他。愿我已无当世望，似君须向古人求。岁寒松柏肯惊秋。

水涵空，山照市。西汉二疏乡里。新白发，旧黄金。故人恩义深。谁

图 3　人工干预后的 11 号类复原图

从复原图可以看出，拼接复原的结果是正确的。经检验，其他类的复原图也是正确的，具体的计算过程略。

3.2.2　英文文件的处理

使用常规方法对英文文件（附件 4 中的碎纸片）处理的效果并不好，拼接复原的正确率只有 60%～70%。这可能是由中、西文文字的差异造成的，如中文是方块字，黑白间隔较为明显，而英文字母则不具备这一特征。例如，a，e，b，h，g，q 这 6 个字母有 3 种特点：a 和 e 只占中间，b 和 h 是上面出头，而 g 和 q 是下面出头。所以用常规方法处理英文文件效果不好。

为提高英文碎纸片聚类的准确率，需要在聚类分析中结合英文字母的特征进行分析，例如选择英文的基线（baseline）作为聚类的依据。图 4 给出了英文基线的定义[4]。

图 4　英文基线的定义图

也可以找到最左侧的 11 块碎片（因为空白处的数值为 255，这一点是能够做到的），再根据基线的特征作聚类分析，将 209 块碎纸片分别对应到 11 个类中。

总之，英文碎纸片的处理要比中文复杂，人工干预节点也较多，而准确率却不如中文的高。

3.3　图形复原

聚类后，对于在同一行的碎片按照问题 1 的方法（TSP 方法）排序，得到 11 个横条碎片文件，重新利用问题 1 的程序对行文件作行排序，得到整个复原图。由于行与行之间存在空白，而且横切面有可能恰好在空白处，所以计算出排序结果很可能不是真正的原始图形，需要根据前后文的意思用手工方式拼接成原始图形。对于中英文文件，其计算结果如下。

中文文件（附件 3）的复原效果较好，复原率能达 100%，并且人工干预较少，在聚类过程中只有前面提到的 1 处人工干预，在恢复成原图形时只需要根据各横行的中文意思排列横行的次序即可。

西文文件（附件 4）的复原效果稍差，人工干预的节点也较多。如果使用常规的聚类方法，再加上人工干预，复原率能达 80%～90% 。如果增加英文基线的特征进行聚类，复原率还会进一步提升。

4　双面信息的使用

问题 3 本质上是重复问题 2 的工作，差别就是每块碎纸片提供了双面信息。如果将一页的双面文件看成两页的单面文件，从表面上看，似乎计算量只增加了一倍，即 209 片碎纸变成 418 片，11 横行的聚类改为 22 横行。但实际上，这种处理方式的计算复杂度会按指数增长，人工干预的数量也随之增加，拼接复原率却会降低。

因此问题 3 求解的关键是双面信息的使用。经研究发现，使用正反面的信息只需要对问题 2 中的方法作两处微小的改动：

1）在作聚类分析时将正反两面数据得到的矩阵相加，再按行求均值，同样是利用向量之间的距离定义两个碎纸片的距离，聚类算法不变。由于使用了正反两面的信息，只需采用常规的聚类方法，即仍然使用 Ward 方法，附件 5 的聚类正确率就可以达到 99%，其结果类似于前面提到的中文聚类，有一类多出 2 块，另外一类少了 2 块，经人工干预后，聚类的正确率达到 100%。

2）在求解旅行商问题时，将正反两面的一块碎纸片看成两块碎纸片，这样 TSP 中的 19 个点变成 38 个，其算法不变。在完成某一横行类的 TSP 计算后，得到 38 个碎纸片的 Hamilton 圈，其结果如下：

143a　200a　086a　187a　131a　056a　138b　045b　137a　061a

094a　098b　121b　038b　030b　042a　084a　153b　186a　186b

153a　084b　042b　030a　038a　121a　098a　094b　061b　137b

045a　138a　056b　131b　187b　086b　200b　143b

从计算结果可以看出，只需从“186a　1S86b”中间断开，就形成一个横条的正反两面，经过完全类似的工作后可得到 22 个横条。对于这 22 个横条，只要按照文中上下文的意思就可以拼接成该文件的正反两面的复原图，也可以对 22 个横条先作 TSP 方法的计算机拼接，再根据上下文意思作手工拼接，这样做的目的是可以减少人工干预的次数与工作强度。

参考文献

[1]360doc. 碎纸复原，真的能做到！[EB/OL]. [2013-11-05]. http://www. 360doc. com/content/12/0320/10/4078497_195874035. shtml.

[2]谢金星，薛毅. 优化建模与 LINDO/LINGO 软件[M]. 北京：清华大学出版社，2005.

[3]谢中华. Matlab 统计分析与应用：40 个案例分析[M]. 北京：北京航天航空大学出版社，2010.

[4]维基百科. 基线的定义[EB/OL]. [2013-11-05]. http://zh. wikipedia. org/wiki/X 字高.

The Mathematical Methods for Reconstruction of Shredded Documents

Xue Yi

(College of Applied Science, Beijing University of Technology, Beijing 100124, China)

Abstract: In this paper, the purely mathematical methods for solving reconstruction of shredded document, problem B of 2013 CUMCM, are presented. The methods are simply summarized as three steps: TSP, cluster analysis and use of two-sided information. And according to the subject requirements, explain the way and time nodes of artificial intervention in the three-step.

Key words: TSP; cluster analysis; stitching of scrapped paper

作者简介

薛毅(1958—)，男，博士，教授，主要研究方向是最优化理论及其应用。