Pure Mathematics 理论数学, 2020, 10(5), 549-555
Published Online May 2020 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2020.105067

Hans汉斯

An Analysis of the Authors of *A Dream of Red Mansions* Based on Equivalence Checking and Feature Clustering

Dongbo Cheng, Xiaoling Ke*, Shixin Lin

College of Mathematics and Data Science, Minjiang University, Fuzhou Fujian
Email: 18359409055@139.com, *Xiaoling_ke1981@163.com

Received: Apr. 27th, 2020; accepted: May 20th, 2020; published: May 27th, 2020

Abstract

The equivalence checking model is introduced to calculate the test statistics *U* and *p* values by selecting the statistical frequency of "red" and "jade". According to *U*-tests and probability comparative table, differences between the first 80 chapters and the last 40 chapters are preliminarily concluded. Many cases are clustered by word frequency with K-means clustering and agglomerative clustering. The results show that there are differences in word frequency used in *A Dream of Red Mansions*, and there is more than one author.

Keywords

Equivalence Checking, Feature Clustering, K-Means Clustering, Agglomerative Clustering

基于等价性检验和特征聚类的《红楼梦》作者分析

程东波，柯小玲*，林施鑫

闽江学院数学与数据科学学院，福建 福州
Email: 18359409055@139.com, *Xiaoling_ke1981@163.com

收稿日期：2020年4月27日；录用日期：2020年5月20日；发布日期：2020年5月27日

摘　要

引入等价性检验模型，选取“红”、“玉”二字统计频数，计算检验统计量*U*与*p*值。根据*U*检验值与概

*通讯作者。

文章引用: 程东波, 柯小玲, 林施鑫. 基于等价性检验和特征聚类的《红楼梦》作者分析[J]. 理论数学, 2020, 10(5): 549-555. DOI: 10.12677/pm.2020.105067

率对照表，初步得出前80章与后40章存在差异，并非一人所著。同时选取K均值聚类与凝聚聚类，根据词频聚类出多种情况。结果表明，《红楼梦》全书使用词频均存在着差异，其作者不止一人。

关键词

等价性检验，特征聚类，K-均值聚类，凝聚聚类

1. 引言

《红楼梦》是中国古典四大名著之一，清代作家曹雪芹创作的章回体长篇小说。由于其在传播和保留过程中出现缺失，当下普遍的观点是前 80 回是曹雪芹所著，高鹗续写后 40 回。

前人的论述已经多从文学用语和数理统计学两大方面反复地论证过，并且得出结论：《红楼梦》前 80 回与后 40 回的作者并非同一人，甚至《红楼梦》全书也非一个作者所著。瑞典汉学家高本汉(B. Karlgren)早在 1952 年用统计学方法分析了 32 个语法和京话与口语词汇的用字习惯，认为全部 120 回均为曹雪芹所著[1]；而胡适先生在 20 世纪 20 年代用文本的文学分析和文献考证方法认定前 80 回为曹雪芹所著，而后 40 回为高鹗所著[2]；至今，更加有人认为，前 80 回也并非出自同一人之手[3]。

采用统计方法进行研究的主要有：李国强[4]等根据《红楼梦》的词频及其相关性进行研究，得到前中后各 40 回的相关度很高，但两两相关度很低的结论。而施建军[5] [6]利用支持向量机方法得到了前 80 回和后 40 回在写作风格上存在明显差别的结论，但其聚类方法不能为判断作者提供可靠的依据。叶雷[7]则使用文体特征进行 K-means 聚类，得到了后 40 回不是前 80 回作者所著的结论。

本文主要利用特征聚类对《红楼梦》前 80 回和后 40 回进行文本分析。分析《红楼梦》120 回的词量、词频，确认其是否为同一作者所著，并用等价性检验模型进行验证。

2. 基于等价性检验的《红楼梦》作者分析模型

2.1. 数据预处理

统计全文中“红”、“玉”的字频，得到表 1 所示数据。根据表 1 的数据可以绘制“红”、“玉”二字在前 80 回和后 40 回的频率，如图 1 所示。由此，容易看出“红”、“玉”两字在前 80 回与后 40 回的频率存在差异。

Table 1. Statistics of frequency of “red” and “jade” in the full text
表 1. 全文中“红”和“玉”的频数统计表

	前 80 回	后 40 回	前 80 回频率	后 40 回频率
“红”	470	153	0.0838%	0.0553%
“玉”	3264	1720	0.5817%	0.6217%
合计	3734	1873	0.6655%	0.6770%
总字数	561,093	276,659		

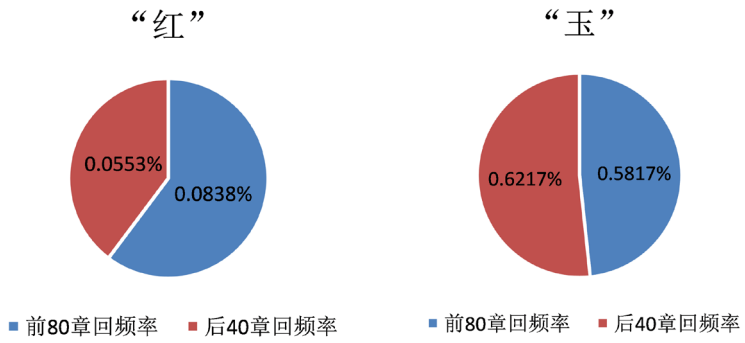

Figure 1. The frequency of the words “red” and “jade” in the first 80 chapters and the latter 40 chapters of *A Dream of Red Mansions*
图 1. 《红楼梦》中“红”“玉”二字在前 80 回频率和后 40 回频率

2.2. 等价性检验模型的建立与求解

选取等价性检验模型，假设前 80 章与后 40 章两个样本并未存在显著差异，即作者为同一人。根据表 1，《红楼梦》前 80 章回和后 40 章回的“红”字和“玉”字出现的频数分析数理统计学的问题，求解等价性检验模型。这一检验问题可化为两个相互独立的二项总体的等价性检验，此时：$X_1 \sim b(n_1, p_1)$ 表示前 80 回的二项分布，其中 $n_1 = 561093$，X_1 表示前 80 章回的“红”字出现的频数，其实测值为 $x_1 = 470$，p_1 表示前 80 章回“红”字出现的频率；$X_2 \sim b(n_2, p_2)$ 表示后 40 回的二项分布，其中 $n_2 = 276659$，X_2 表示后 40 章回的“红”字出现的频数，其实测值为 $x_2 = 153$，p_2 表示后 40 章回“红”字出现的频率。

采用渐进正态检验，计算检验统计量 U，进而计算检验的 p 值，即：否定原假设而犯错误的概率。解得记“红”字的检验统计量 $U_1 \approx 4.494465$。重复上述步骤解得“玉”字的检验统计量 $U_2 \approx 2.238018$。等价性检验模型的 U 检验值与概率对照表如表 2 所示。

Table 2. Comparison between U test value and probability
表 2. U 检验值与概率对照表

U 检验值	p-值	可信概率%
2.4360	0.0074	99.26%
2.3590	0.0092	99.08%
1.8622	0.0313	96.87%
1.5811	0.0569	94.31%
1.4325	0.0760	92.40%

通过表 2 可得：U 检验值越大，则可信概率就越大，即：拒绝原假设，差异就越显著。从数值模拟获得的数据分析，“红”、“玉”两字的 U 检验量足够大，表明前 80 回与后 40 回确实存在显著差异。但无论观察图表，或是对照等价性检验的结论，都可以发现其中明显的问题：等价性检验所呈现的结果存在偏差，寻找其原因也能很容易发现，“玉”字与主角名字有很大的关系，而主角之一的“林黛玉”在后 40 章已经去世，这就对这个字的频数造成一定的影响。

综上，该方法中存在两个缺陷。其一，适用性差。因为需要抽取样本逐一进行等价性检验，对多个样本重复操作的过程将会十分繁琐；其二，模型稳定性差。针对于“红”和“玉”两个样本，出现了 U 检验值不小的差异，足以说明不同样本之间存在的差异性会受到其余因素的影响，且每个样本之间可能会出现矛盾。

所以，进一步选取特征聚类验证继续验证猜想。

3. 基于聚类分析的《红楼梦》作者分析

3.1. 数据处理

以每十章回为样本，分别对特征词频进行统计，选取代词进行统计，结果如图 2。

```
{'你': 340, '此书': 2, '此': 60, '自己': 37, '这': 316, '那': 129,
{'你': 410, '我': 476, '这里': 53, '有人': 12, '各处': 19, '什么': 8
{'那里': 53, '她': 233, '这': 269, '我': 849, '你': 664, '你们': 58
{'自己': 63, '她': 315, '你': 570, '怎么样': 15, '怎么': 75, '他': 2
{'我': 902, '这': 349, '那': 155, '他': 449, '它': 13, '你': 692, '
{'各': 37, '这': 442, '其一': 1, '其二': 1, '其三': 1, '汝': 1, '他
{'那': 145, '你': 536, '什么': 119, '我': 689, '这': 331, '这一': 1
{'诸事': 1, '在家': 8, '诸': 4, '之物': 9, '本': 10, '有人': 32, '别
{'这': 260, '自己': 97, '他': 239, '你': 489, '为什么': 24, '这样':
{'她们': 17, '什么': 129, '自己': 85, '那': 132, '如何': 17, '为什么
{'她': 228, '有人': 10, '什么': 152, '那': 128, '那些': 46, '我': 62
{'自己': 79, '她': 251, '这时': 2, '别的': 12, '这晚': 1, '我': 601,
```

Figure 2. Example of word frequency after word segmentation by part of speech (part of speech)
图 2. 按照词性(词性)分词后统计的词频示例

根据分词结果统计每十回的代词词量，如下表 3 所示。

Table 3. Number of pronouns per decade
表 3. 每十回的代词词量

1~10 回	11~20 回	21~30 回	31~40 回	41~50 回	51~60 回	61~70 回	71~80 回	81~90 回	91~100 回	101~110 回	111~120 回
218	188	158	169	155	187	179	174	139	148	148	163

其次，将各个代词的词频按照递减的顺序排列，选取词频排行前 50 的代词词频来作为第一次选取的特征向量，如图 3。

```
[('我', 420), ('你', 340), ('这', 316), ('他', 309), ('她', 161), ('那', 129), ('什么', 93), (
[('我', 476), ('你', 410), ('这', 199), ('他', 197), ('她', 158), ('那', 130), ('什么', 88), (
[('我', 849), ('你', 664), ('他', 339), ('这', 269), ('她', 233), ('那', 164), ('什么', 159),
[('我', 761), ('你', 570), ('她', 315), ('这', 307), ('他', 234), ('那', 169), ('什么', 163),
[('我', 902), ('你', 692), ('他', 449), ('这', 349), ('她', 201), ('那', 155), ('什么', 115),
[('我', 699), ('你', 644), ('她', 491), ('这', 442), ('他', 263), ('我们', 167), ('那', 150),
[('我', 689), ('你', 536), ('她', 392), ('这', 331), ('他', 325), ('那', 145), ('我们', 141),
[('我', 757), ('你', 545), ('这', 449), ('她', 411), ('他', 199), ('什么', 151), ('我们', 128)
[('我', 553), ('你', 489), ('这', 260), ('他', 239), ('她', 173), ('什么', 172), ('那', 161),
[('我', 525), ('你', 394), ('他', 345), ('她', 221), ('这', 196), ('那', 132), ('什么', 129),
[('我', 628), ('你', 368), ('他', 239), ('她', 228), ('这', 220), ('什么', 152), ('那', 128),
[('我', 601), ('你', 395), ('他', 262), ('那', 257), ('她', 251), ('这', 199), ('什么', 145),
```

Figure 3. First eigenvector
图 3. 第一特征向量

显然在第一特征向量的代词中，并不是每个样本的第一特征向量的元素都相同，因此进一步提取数据。在第一特征向量中，统计在每个样本中都出现的元素，如图 4。

根据我们划分样本的方式，上图中的数字为 12 的即表示每个样本中都包含的元素。因此，将每个样本中都包含的元素作为第二特征向量，每个向量中包含 27 个元素，即为最终选取的特征向量。并将每个样本中的数据进行提取，最终的数据提取的结果如图 5 所示。

{'我': 12, '你': 12, '这': 12, '他': 12, '她': 12, '那': 12, '什么': 12, '其': 11, '这里': 12, '我们': 12, '此': 9, '那里': 12, '谁': 12, '自己': 12, '这个': 12, '他们': 12, '你们': 12, '咱们': 12, '这样': 12, '怎么': 12, '如何': 11, '有些': 12, '这些': 12, '那边': 12, '在家': 9, '这么': 12, '那些': 12, '有人': 12, '该': 12, '何': 5, '这边': 10, '本': 6, '各': 10, '这话': 12, '有个': 9, '其中': 1, '吾': 1, '这日': 3, '别人': 12, '一样': 9, '那日': 7, '另': 8, '别的': 11, '哪里': 10, '每': 2, '有事': 4, '他家': 1, '彼': 1, '怎': 2, '那个': 9, '她们': 11, '各处': 8, '怎么样': 8, '诸': 1, '此处': 1, '何处': 1, '彼时': 2, '那时': 9, '别': 11, '哪': 5, '每日': 5, '之中': 1, '为什么': 8, '前儿': 7, '这么着': 4, '彼此': 1, '谁家': 2, '各自': 7, '各色': 2, '这般': 4, '这一': 2, '每人': 1, '那样': 2, '你老': 1, '之物': 1, '此事': 2, '这事': 2, '何等': 1, '怎样': 3, '别处': 2, '然': 1, '这种': 2, '爷们': 3, '那块': 1, '有时': 1, '我家': 2, '怎么着': 1, '那年': 1}

Figure 4. Elements in each sample
图 4. 每个样本中包含的元素情况

[('我', 420), ('你', 340), ('这', 316), ('他', 309), ('她', 161), ('那', 129), ('什么', 93), (
[('我', 476), ('你', 410), ('这', 199), ('他', 197), ('她', 158), ('那', 130), ('什么', 88), (
[('我', 849), ('你', 664), ('他', 339), ('这', 269), ('她', 233), ('那', 164), ('什么', 159),
[('我', 761), ('你', 570), ('她', 315), ('这', 307), ('他', 234), ('那', 169), ('什么', 163),
[('我', 902), ('你', 692), ('他', 449), ('这', 349), ('她', 201), ('那', 155), ('什么', 115),
[('我', 699), ('你', 644), ('她', 491), ('这', 442), ('他', 263), ('我们', 167), ('那', 150),
[('我', 689), ('你', 536), ('她', 392), ('这', 331), ('他', 325), ('那', 145), ('我们', 141),
[('我', 757), ('你', 545), ('这', 449), ('她', 411), ('他', 199), ('什么', 151), ('我们', 128)
[('我', 553), ('你', 489), ('这', 260), ('他', 239), ('她', 173), ('什么', 172), ('那', 161),
[('我', 525), ('你', 394), ('他', 345), ('她', 221), ('这', 196), ('那', 132), ('什么', 129),
[('我', 628), ('你', 368), ('他', 239), ('她', 228), ('这', 220), ('什么', 152), ('那', 128),
[('我', 601), ('你', 395), ('他', 262), ('那', 257), ('她', 251), ('这', 199), ('什么', 145),

Figure 5. Finally selected data example
图 5. 最终选取的数据实例

根据选取的 27 个特征值，依次提取相应的词频，组成特征向量，如图 6。

[[420 340 316 309 161 129 93 67 61 56 39 37 37 35 34 33 31 29
   25 23 21 19 17 17 16 12 12]
 [476 410 199 197 158 130 88 53 63 26 50 50 39 27 46 33 35 34
   10 32 14 20 15 12 24 16 22]
 [849 664 269 339 233 164 159 66 57 53 65 83 84 33 58 22 27 82
   13 47 25 40 20 23 28 40 36]
 [761 570 307 234 315 169 163 55 112 45 57 63 99 33 89 41 69 75
    9 47 17 47 35 21 23 49 28]
 [902 692 349 449 201 155 115 70 103 75 40 56 65 66 114 51 54 52
   10 38 21 40 14 10 29 24 16]
 [699 644 442 263 491 150 135 92 167 39 64 52 68 64 148 73 45 50
   18 61 10 27 20 25 27 38 20]
 [689 536 331 325 392 145 119 81 141 56 61 112 65 32 78 78 55 61
   13 39 44 11 12 15 35 27 21]
 [757 545 449 199 411 117 151 67 128 39 103 63 62 45 114 62 104 58
   20 51 36 11 14 32 35 42 31]
 [553 489 260 239 173 161 172 120 76 126 45 97 55 46 90 44 64 101
   21 36 46 35 18 15 17 27 19]
 [525 394 196 345 221 132 129 77 78 123 36 85 47 32 60 58 40 74
   19 45 37 43 47 15 29 21 11]

Figure 6. 12 eigenvector data of samples
图 6. 12 个样本的特征向量数据

3.2. K 均值聚类和凝聚聚类的聚类结果分析

使用 python 机器学习模块 Scikit-Learn 模块进行求解。

将 K 均值聚类模型实例化，将处理后的特征向量应用到算法模型中，解得：

```
2 cluster:
 [1 1 0 0 0 0 0 0 1 1 1 1]
```

Figure 7. K-means clustering of two categories
图 7. K 均值聚类两个类别的聚类情况

从图 7 来看，容易发现第 21 回到第 80 回分自然的分在了一个类别，而前 20 回却和后 40 回分在了一起，这说明：前 20 回合后 40 回有更好的相似性。这不符合预计结果，猜想可能原因是前 80 回内部存在相似性存在一定的差异。

为进一步分析上述结论的原因，现将特征向量应用到新的算法模型中，该算法模型训练 3 个簇，即 3 个类别。实验结果如下图 8 所示：

```
3 cluster:
 [0 0 1 2 1 2 2 2 0 0 0 0]
```

Figure 8. K-means clustering of three categories
图 8. K 均值聚类三个类别的聚类情况

从图 9 的实验结果来看，我们很容易发现从第 21 回到 80 回内部存在一定的差异性，被分成了两个类别；而前 20 回和后 40 回仍被分在了一个类别。该实验结果只说明了：第 21 回到第 80 回内部存在差异性。

再次训练新的算法模型，该算法模型训练 4 个类别，实验结果如图 9。

```
4 cluster:
 [2 2 3 1 3 1 1 1 0 0 0 0]
```

Figure 9. K-means clustering of four categories
图 9. K 均值聚类四个类别的聚类情况

这个实验结果很明显看出，前 80 回合后 40 回内容存在差异，不是同一个人所著。但同时，前 80 回同样存在差异，因此我们可以以这个结果进一步进行猜想，《红楼梦》不仅有两个作者，可能存在两个以上的编撰者。

凝聚聚类算法的处理方式与 K 均值聚类的方式相似，将图 6 的特征向量应用到凝聚聚类算法模型中。分别试验聚成两个类、三个类、四个类的算法模型，实验结果如图 10。

```
2 cluster:
 [0 0 1 1 1 1 1 1 0 0 0 0]
3 cluster:
 [2 2 0 0 0 0 0 0 1 1 1 1]
4 cluster:
 [2 2 0 0 0 3 0 3 1 1 1 1]
```

Figure 10. Clustering of clustering into 2, 3 and 4 categories
图 10. 凝聚聚类分别聚成 2，3，4 个类别的聚类情况

从图 10 很容易发现：在采用凝聚聚类聚成 3 个类别的时候，能够很显然的体现出前 80 回和后 40 回作者不一，而在前 80 回中，前 20 回和其他 60 回作者存在差异。

4. 总结与展望

综合上述模型求解结果，等价性检验模型虽然作为数理统计中优秀的模型，但依旧有着不小的局限性，在处理多个样本的情况下，K 均值聚类和凝聚聚类反而得到了优秀的结果。从最终的实验结果可以观察出，根据分析《红楼梦》前 80 回和后 40 回的部分词量和词频，可以表明《红楼梦》前 80 回和后

40 回并不是同一作者所著。除此之外，我们还发现，在前 80 回内的文章也存在着不小的差异。因此可以一定程度上表明，《红楼梦》不止一个作者所著。

基金项目

闽江学院校长基金(103952018230)。

参考文献

[1] 胡适.《红楼梦考证》(改定稿) [M]. 北京: 北京出版社, 2015.

[2] Karlgren, B. (1952) New Excursions in Chinese Grammar. *The Bulletin of the Museum of Far Eastern Antiquities*, **24**, 79.

[3] Koppel, M., Schler, J. and Argamon, S. (2009) Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, **60**, 9-26. https://doi.org/10.1002/asi.20961

[4] 李国强, 李瑞芳. 基于计算机的词频统计研究——考证《红楼梦》作者是否唯一[J]. 沈阳化工学院学报, 2006, 20(4): 305-307.

[5] 施建军. 基于支持向量机技术的《红楼梦》作者研究[J]. 红楼梦学刊, 2011(5): 35-52.

[6] 施建军. 关于以《红楼梦》120 回为样本进行其作者聚类分析的可信度问题研究[J]. 红楼梦学刊, 2010(5): 318-335.

[7] 叶雷. 基于计量文体特征聚类的《红楼梦》作者分析[J]. 红楼梦学刊, 2016(5): 312-324.