【数据挖掘实验】聚类分析方法

原创

想飞的蓝笨笨 于 2021-01-27 20:17:26 发布 2029 收藏 3

分类专栏： Clemetine

本文链接：https://blog.csdn.net/qq_44762986/article/details/113269207

一、实验项目名称：

聚类分析方法

二、实验目的与要求：

在软件方面：会用Clementine软件进行聚类分析。

在理论方面：聚类分析及其常用的聚类分析方法，数据挖掘中的聚类分析。

三、实验原理：

1、聚类分析方法

聚类分析是数据分析中的一种重要技术，它的应用极为广泛。许多领域中都会涉及聚类分析方法的应用与研究工作。例如：在科学数据探测、信息检索、文本挖掘、空间数据库分析、Web数据分析、客户关系管理、医学诊断、生物学等方面的数据挖掘应用软件中，聚类分析技术都起着重要作用。在商业领域，聚类可以帮助市场分析人员从消费者数据库中分出不同的消费群体来，并且概括出每一类消费者的消费模式或者说习惯，发现不同类型的客户群，可以用来分类具有相似功能的基因，了解种群的内在结构。聚类还可以用来从地理数据库中识别出具有相似土地用途的区域；可以从保险公司的数据库中发现汽车保险中具有较高索赔概率的群体；还可以从一个城市的房地产信息数据库中，根据户型、房价及地理位置将房地产分成不同的类；还可以用来对Web上不同类型的文档进行分类等。

我们主要讲的方法是谱系聚类、快速聚类、两步聚类。

2、聚类分析方法应用

聚类分析在《红楼梦》作者问题上的应用

众所周知,《红楼梦》一书共120回,自从胡适作《红楼梦考证》以来,一般都认为前80回为曹雪芹所写,后40回为高鹗所续。然而长期以来这种看法一直都饱受争议。能否从统计上做出论证从1985年开始,复旦大学的李贤平教授带领他的学生作了这项很有意义的工作,他们创造性的想法是将120回看成是120个样本,然后确定与情节无关的虚词出现的次数作为变量,巧妙运用数理统计分析方法,看看哪些回目出自同一人的手笔。一般认为,每个人使用某些词的习惯是特有的。于是李教授用每个回目中47个虚词(之,其,或,亦…,呀,吗,咧,罢……可,便,就……等)出现的次数(频率),作为《红楼梦》各个回目的数字标志。之所以要抛开情节,是因为在一般情况下,同一情节大家描述的都差不多,但由于个人写作特点和习惯的不同,所用的虚词是不会一样的。利用多元分析中的聚类分析法进行聚类,果然将120回分成两类,即前80回为一类,后40回为一类,很形象地证实了不是出自同一人的手笔。之后又进一步分析前80回是否为曹雪芹所写这时又找了一本曹雪芹的其它著作,做了类似计算,结果证实了用词手法完全相同,断定前80回为曹雪芹一人手笔,是他根据《石头记》写成,中间插入《风月宝鉴》,还有一些别的增加成分。而后40回是否为高鹗写的呢 论证结果推翻了后40回是高鹗一个人所写,而是曹雪芹亲友将其草稿整理而成,宝黛故事为一人所写,贾府衰败情景当为另一人所写等等。这个论证在红学界轰动很大,李教授他们用多元统计分析方法支持了红学界的观点,红学界大为赞叹[11]。

所谓聚类分析，顾名思义，就是按照某种标准将样本物以类聚。即使续作者刻意模仿作者的写法，但是文风是不能模仿的，而对虚词的使用是难以做到一致的，这就是标准（也就是统计量）所在。李教授的工作便是证明了 前八十回和后四十回在虚词的使用上截然不同。而石头记与风月宝鉴的对比使用的则是因子分析的方法。每一回四十七个虚词出现不同次数，而一共有120回，这样就构成一个47*120的矩阵，李教授在统计软件SPSS上分析这个大型矩阵得到以上结果，可信度甚高，因为它是完全客观不带有主观色彩的方法，仅从文本入手。就凭这一点，比某些胡说八道的红学家强之百倍。

四、实验方案设计：

数据源背景分析；选择聚类方法；分析聚类结果。

五、测试数据与实验结果

测试数据1：谱系聚类

中国男足可算是杯具到家了，几乎到了过街老鼠人人喊打的地步。对于目前中国男足在亚洲的地位，各方也是各执一词，有人说中国男足亚洲二流，有人说三流，还有人说根本不入流，更有人说其实不比日韩差多少，是亚洲一流。既然争论不能解决问题，我们就让数据告诉我们结果吧。下图是采集的亚洲15只球队在2005年-2010年间大型杯赛的战绩（由于澳大利亚是后来加入亚足联的，所以这里没有收录）。

	A	B	C	D
1		2006年世界杯	2010年世界杯	2007年亚洲杯
2	中国	50	50	9
3	日本	28	9	4
4	韩国	17	15	3
5	伊朗	25	40	5
6	沙特	28	40	2
7	伊拉克	50	50	1
8	卡塔尔	50	40	9
9	阿联酋	50	40	9
10	乌兹别克斯坦	40	40	5
11	泰国	50	50	9
12	越南	50	50	5
13	阿曼	50	50	9
14	巴林	40	40	9
15	朝鲜	40	32	17
16	印尼	50	50	9

其中包括两次世界杯和一次亚洲杯。提前对数据做了如下预处理：对于世界杯，进入决赛圈则取其最终排名，没有进入决赛圈的，打入预选赛十强赛赋予40，预选赛小组未出线的赋予50。对于亚洲杯，前四名取其排名，八强赋予5，十六强赋予9，预选赛没出现的赋予17。这样做是为了使得所有数据变为标量，便于后续聚类。下面先对数据进行[0,1]规格化，下面是规格化后的数据：

	A	B	C	D
1		2006年世界杯	2010年世界杯	2007年亚洲杯
2	中国	1	1	0.5
3	日本	0.3	0	0.19
4	韩国	0	0.15	0.13
5	伊朗	0.24	0.76	0.25
6	沙特	0.3	0.76	0.06
7	伊拉克	1	1	0
8	卡塔尔	1	0.76	0.5
9	阿联酋	1	0.76	0.5
10	乌兹别克斯坦	0.7	0.76	0.25
11	泰国	1	1	0.5
12	越南	1	1	0.25
13	阿曼	1	1	0.5
14	巴林	0.7	0.76	0.5
15	朝鲜	0.7	0.68	1
16	印尼	1	1	0.5

请用谱系聚类（SPSS软件）对上述表格中的数据进行聚类，每一年都聚为3类，观察中国在这3年中的同类成员是否有变化。

实验结果2：

（1）对2006年世界杯进行谱系聚类，谱系图及聚类结果如下：

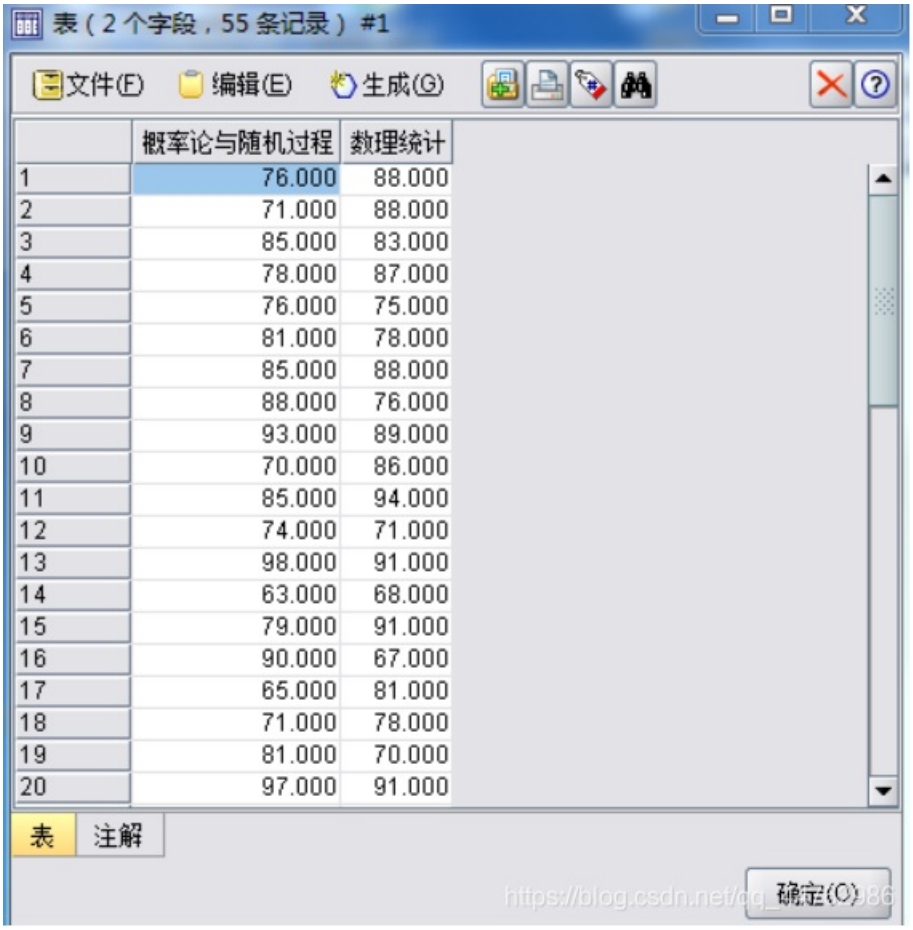
（2）对2010年世界杯进行谱系聚类，谱系图及聚类结果如下：

（3）对2007年亚洲杯进行谱系聚类，谱系图及聚类结果如下：

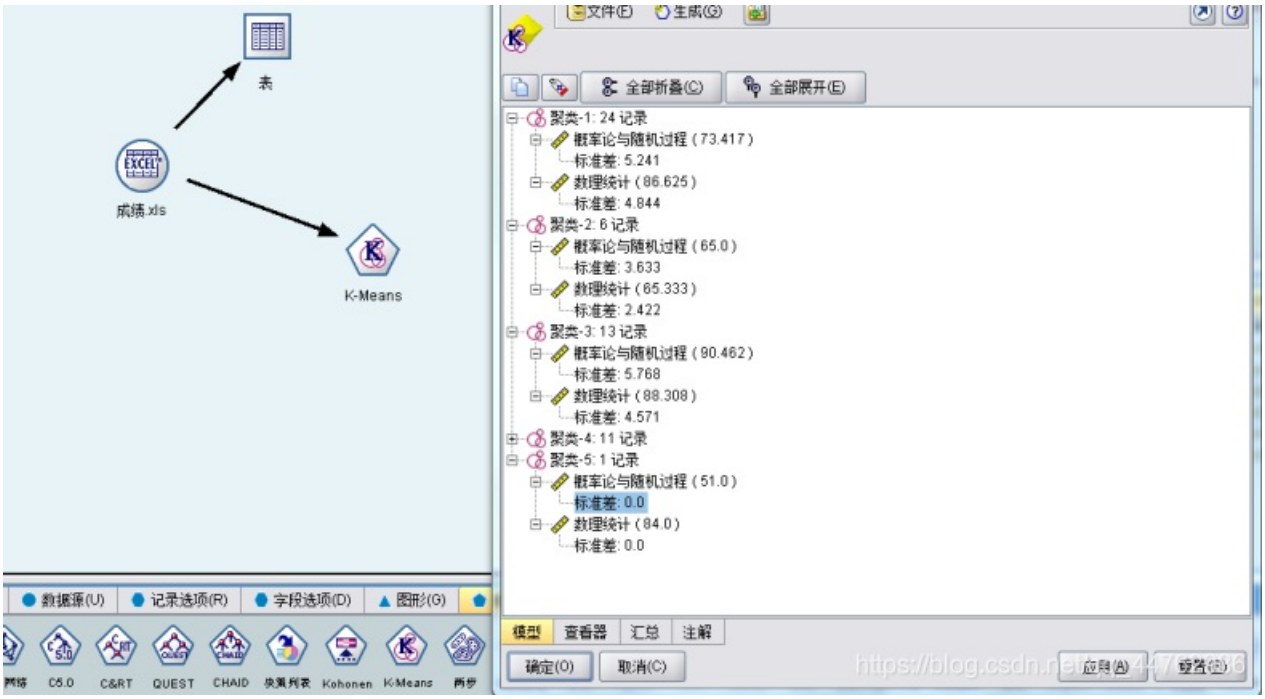
测试数据2：快速聚类

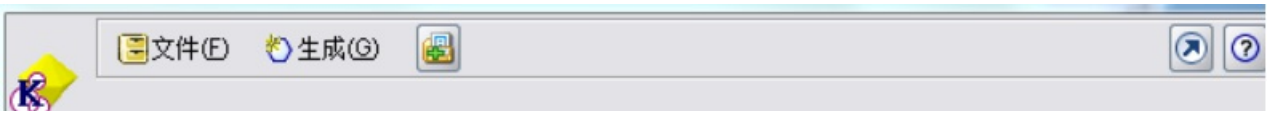
以附录中的成绩为数据源，用SPSS或者Clementine对该数据源是本班学生的概率论与随机过程和数理统计两门课程的成绩，通过快速聚类，将其聚为3类或5类，给出每一个人所属类别，并给出类中心，观查每一类的特点。

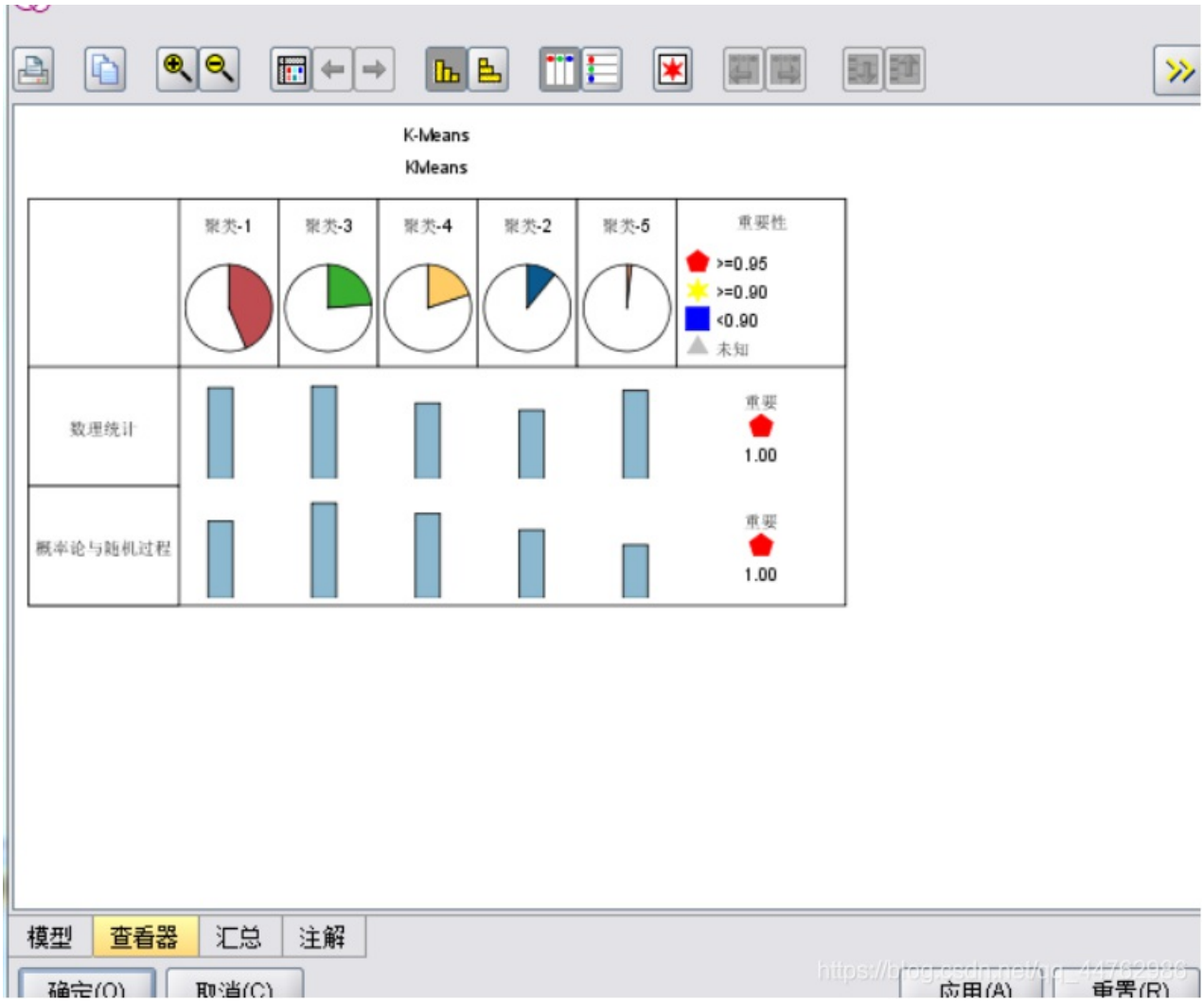
实验结果



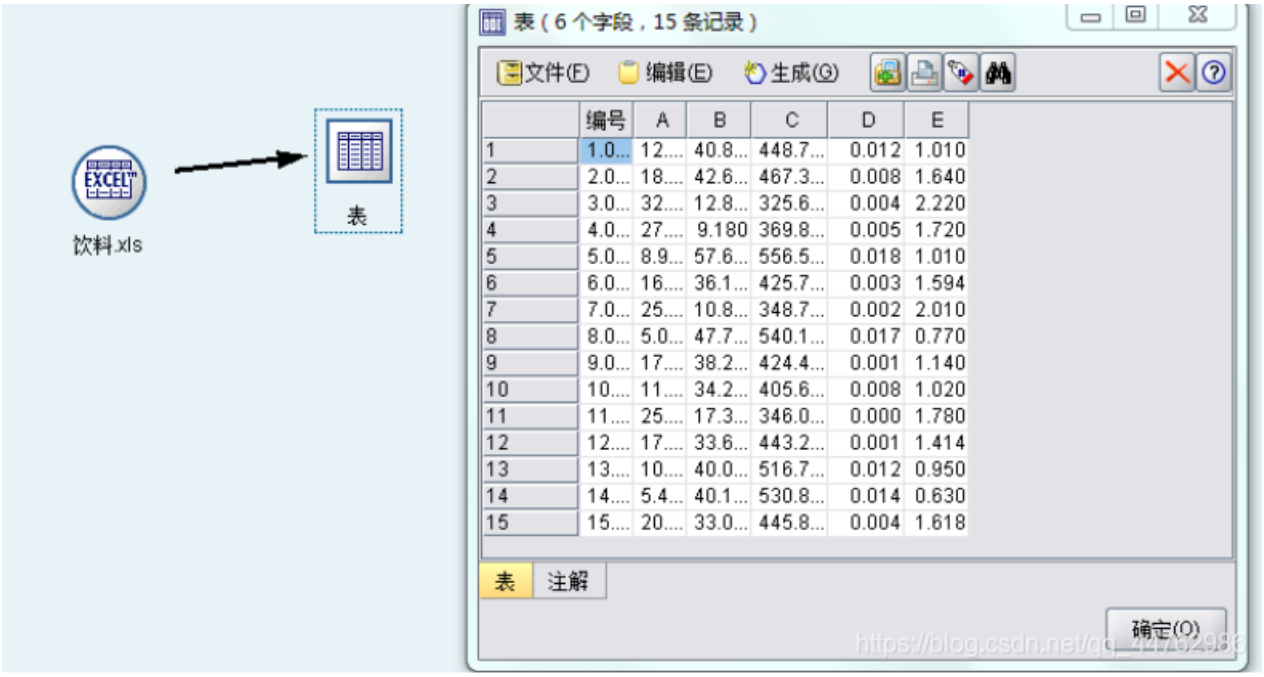
	概率论与随机过程	数理统计
1	76.000	88.000
2	71.000	88.000
3	85.000	83.000
4	78.000	87.000
5	76.000	75.000
6	81.000	78.000
7	85.000	88.000
8	88.000	76.000
9	93.000	89.000
10	70.000	86.000
11	85.000	94.000
12	74.000	71.000
13	98.000	91.000
14	63.000	68.000
15	79.000	91.000
16	90.000	67.000
17	65.000	81.000
18	71.000	78.000
19	81.000	70.000
20	97.000	91.000

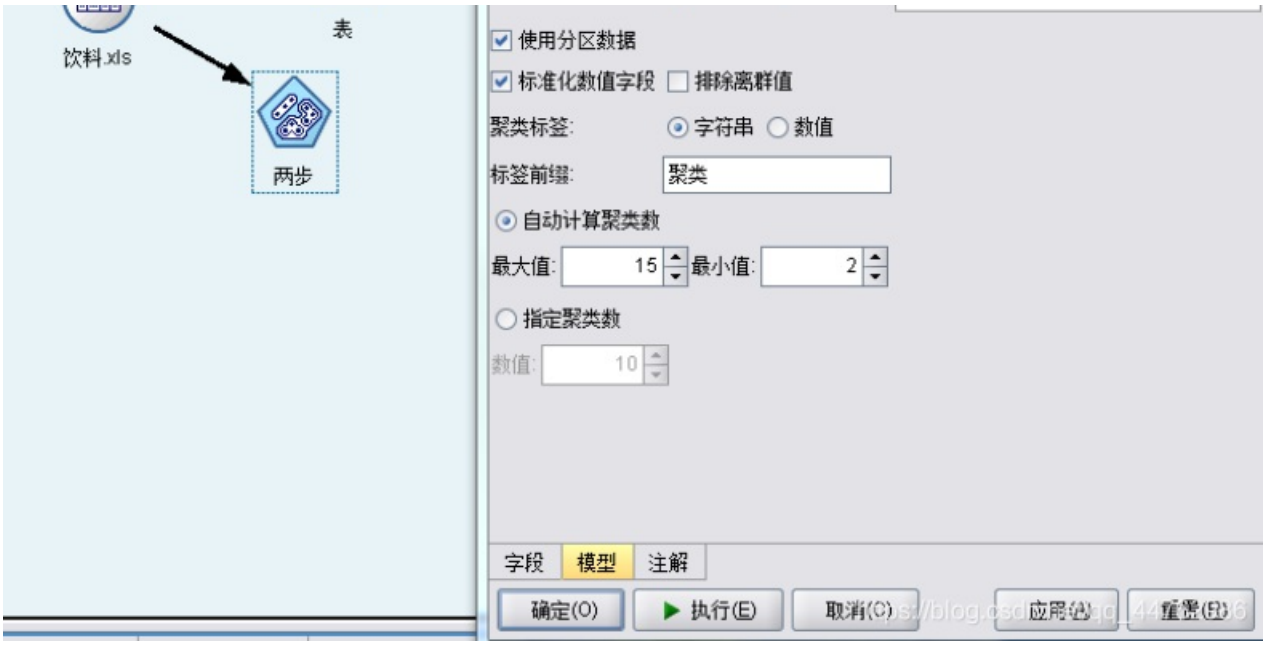






测试数据3：用SPSS或者Clementine，利用两步聚类完成教材中的示例（教材72页），重点分析74页的结果。
实验结果：





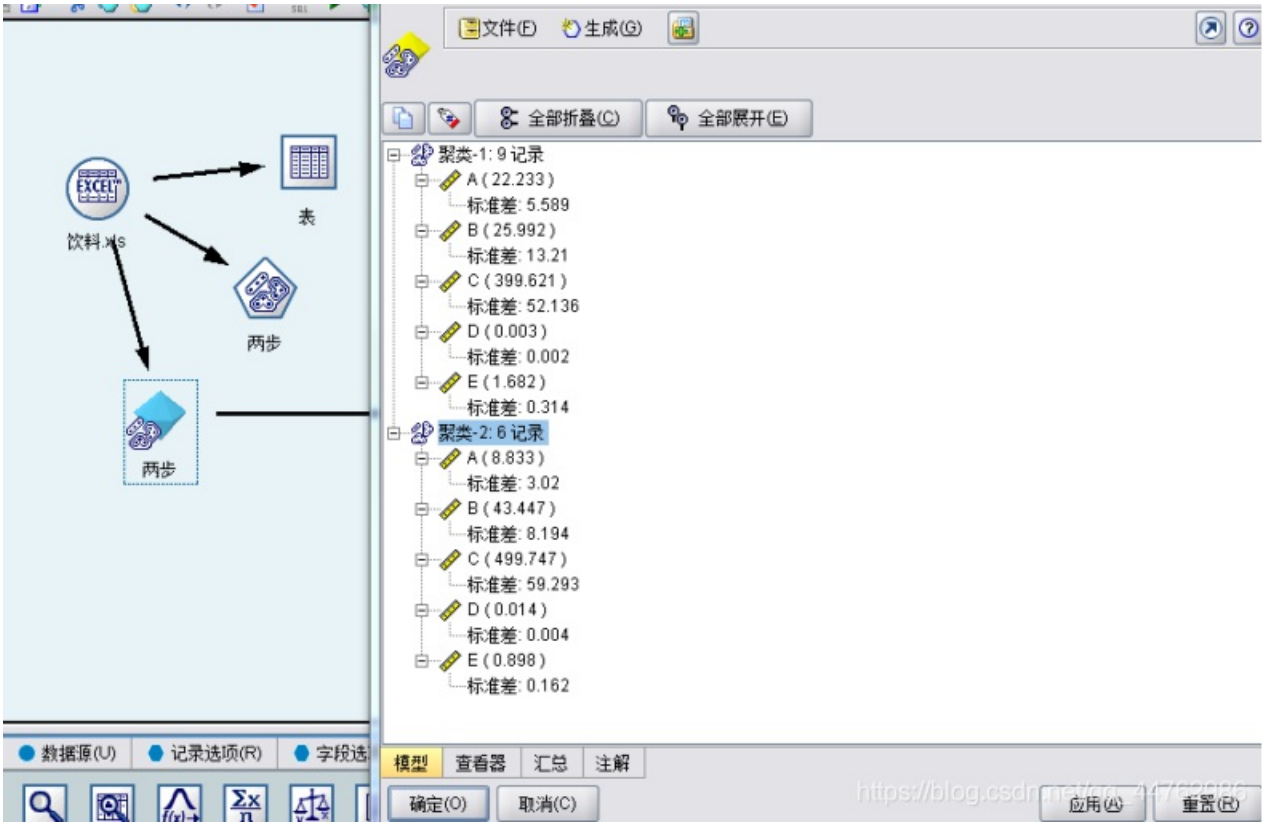


表 (7 个字段，15 条记录)

文件(F) 编辑(E) 生成(G)

	编号	A	B	C	D	E	$T-两步
1	1.0...	12....	40.8...	448.7...	0.012	1.010	聚类-2
2	2.0...	18....	42.6...	467.3...	0.008	1.640	聚类-1
3	3.0...	32....	12.8...	325.6...	0.004	2.220	聚类-1
4	4.0...	27....	9.180	369.8...	0.005	1.720	聚类-1
5	5.0...	8.9...	57.6...	556.5...	0.018	1.010	聚类-2
6	6.0...	16....	36.1...	425.7...	0.003	1.594	聚类-1
7	7.0...	25....	10.8...	348.7...	0.002	2.010	聚类-1
8	8.0...	5.0...	47.7...	540.1...	0.017	0.770	聚类-2
9	9.0...	17....	38.2...	424.4...	0.001	1.140	聚类-1
10	10....	11....	34.2...	405.6...	0.008	1.020	聚类-2
11	11....	25....	17.3...	346.0...	0.000	1.780	聚类-1
12	12....	17....	33.6...	443.2...	0.001	1.414	聚类-1

六、实验总结

七、部分参考代码（可附页或提交电子版）

附录：测试数据2的数据源"本班两门课的成绩"

76 88
71 88
85 83
78 87
76 75
81 78
85 88
88 76
93 89
70 86
85 94
74 71
98 91
63 68
79 91
90 67
65 81
71 78
81 70
97 91
70 65
77 95
61 67
63 67
78 75
86 87
77 88
84 83
64 88
65 82
77 72
100 98
96 87
83 71
92 84
71 86
75 80
71 92
88 90
81 83
51 84

69 87
69 62
74 95
64 63
71 80
76 88
79 95
78 71
87 83
82 66
82 84
73 82
67 85
80 90