基于聚类与判别分析法对明星微博进行分类

小组成员：李怡璠 201601101　宋朋燕 201602108

顾赵凯鹭 201603005　范颖斐 201656002　李佳雯 201656011

摘要：本小组旨在定位明星们在流量经济时代中所扮演的角色，借助 SPSS 软件，运用多元统计方法——聚类分析与判别分析，对明星的微博数据，根据其受欢迎程度与活跃程度进行分类分析，并得出以下结论：1）由聚类分析得，样本可分为四类：“大明星”“流量王”“艺术家”和“普通艺人”；2）判别函数回判正确率为 94.5%，对 5 个新样本判别较好，仅 1 个不符合预期。

一、　课题背景：

在移动互联网时代，随着智能手机的普及，在线社交网络已经逐步融入人们的日常生活。第 40 次《中国互联网络发展状况统计报告》数据显示，截至 2017 年 6 月，中国网民规模来到 7.51 亿，互联网普及率为 54.3%；手机网民规模达 7.24 亿，移动互联网已渗透到人们生活的方方面面。在社交网络发展的同时，娱乐圈作为大众关注的重心，也依托此类新媒体逐渐透明化，互动化，大众化。在“热搜当道”与“全民围脖”的时代，在微博中引起广泛关注的不仅有演员艺人、娱乐博主，还有工作室与一些热点节目的官方代理人，无论是草根网红还是明星大 V，每个人都有可能成为流量中心。故评价一个明星的微博影响力，不仅取决于其在相应领域的受欢迎程度，同样也受到其微博内容的吸引力及与粉丝的互动程度，而研究一个明星的影响力有助于我们更好地定位该明星在粉丝经济时代的“吸金能力”。本文即借此兴趣，试图运用数据处理的技术与多元统计的方法，根据微博中不同明星的受欢迎程度与活跃程度进行分类分析。

二、　数据选取：

新浪微博是汇聚明星最全，用户群体最广的站点，我们根据新浪微博中筛选部分明星微博（包括艺人、运动员、网红等），取得的 1747 条数据，进行处理分析。

被解释变量：

微博账号：文本格式，为明星微博的账号名称，样本选取的个体均为微博认证用户（包括个人实名认证的橙色用户与企业实名认证的蓝色用户），且其微博名称不会

经常修改，故可选用此变量作为辨别变量。

解释变量：

关注：连续变量（单位：个），取值范围[0，4833]，是明星关注其他微博的个数，可用于反映该明星的微博特点。

微博数：连续变量（单位：个），取值范围[2，193299]，为该明星截至数据采集时间时仍保留的发微博的个数，可用于反映该明星的微博活跃程度。

微博等级：连续变量（单位：等级 Lv）,取值范围[2，46]，根据用户活跃天数确定，等级高的用户可以享有更多的特权，可用于反映该明星的微博活跃程度。

粉丝数：连续变量（单位：人），取值范围[2239，90928261]，可用于反映明星的受欢迎程度。

微博会员：连续变量，取值范围[1，6]，明星注册微博的会员等级，需主动充值，可以享有更多的特权，在一定程度上可反映该明星的影响力。

爱慕值：连续变量（单位：分），取值范围[22,5626047]，为粉丝对该用户的关注与贡献度，通过粉丝俱乐部签到或者向该明星用户送花送礼物可以提高爱慕值。该值可反映该明星的受欢迎程度与影响力。

收到花数：连续变量（单位：束），取值范围[11,2763629]，为粉丝对该用户送花/礼物的数量，花/礼物一般需要购买，故该值可以反映该明星的受欢迎程度与影响力。

三、 数据处理：

1. 数据清洗

原始数据较为粗糙，我们主要进行了以下简单的清洗工作：

（一）删除重复数据：使用 SPPS 标记相同微博名称的观测值并删除重复信息；

（二）处理缺失值：我们将存在缺失值的数据进行主观分类，对于粉丝数相对较大的不完整数据进行人工填写，对于影响不大的数据进行简单删除。

2. 变量选取

考虑到数据中的变量可能存在重叠信息，故对于清洗后的数据，我们通过 SPSS 筛选参与分类的变量。首先，我们选用主成分分析的方法对除微博账号以外的 7 项变量进行因子分析，分析的结果如下表所示：

成份矩阵[a]

	成份	
	1	2
Follow	.626	-.334
Follower	.596	.144
Weibo_number	.678	-.342
Weibo_level	.812	-.145
Likes	.228	.953
Flowers	.230	.953
VIP	.617	.059

提取方法 :主成份。

a. 已提取了 2 个成份。

解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	2.353	33.608	33.608	2.353	33.608	33.608
2	2.091	29.878	63.486	2.091	29.878	63.486
3	.976	13.943	77.429			
4	.783	11.189	88.618			
5	.428	6.110	94.727			
6	.368	5.251	99.978			
7	.002	.022	100.000			

提取方法：主成份分析。

成份矩阵建议我们提取 2 个主成分，但解释的总方差表却显示提取成分数为 2 时累积解释的总方差仅有 63.486%，并不能很好地体现原始数据的重要信息，即所得的各个主成分浓缩原始变量信息的能力差别不大，故我们不选择用主成分分析的方法进行降维。

除此之外，为验证上述结论与在一定程度上剔除重叠信息，我们对变量进行相关性检验，发现绝大多数变量相关性小于 0.3，但爱慕值与收到花数之间存在着极强的相关性，Pearson 相关系数高达 0.998，故从数据检验与实际意义的判断（粉丝向明星送花是增加明星爱慕值的途径之一），我们可以认为收到花数与爱慕值反映同一信息，故在具体聚类与判别分析时我们删除收到花数的变量，仅选用爱慕值。

四、 聚类分析

1. 系统聚类法

1.1 确定聚类距离测算方法

为确定聚类时组内距离测算、组间距离测算等细节，我们对样本数据进行描述统计。

表 1 样本描述统计

描述统计					
	N	最小值	最大值	均值	标准 偏差
关注	490	0	2413	325.97	246.993
粉丝	490	37109	90928261	10613090.52	15549923.950
微博数	490	62	9328	2053.48	1847.193
微博等级	490	4	46	33.06	7.219
爱慕值	490	22	5626047	121054.57	474047.055
微博会员	490	1	6	5.52	.911
有效个案数	490				

描述统计显示，各变量在数值上相差巨大。为避免因为数量级差异而对聚类结果造成的影响，我们先利用 Z 得分对变量进行标准化处理，再聚类分析。标准化后变量的描述统计如下：

表 2 样本标准化后的描述统计

描述统计					
	N	最小值	最大值	均值	标准 偏差
Zscore(关注)	490	-1.31976	8.44976	.0000000	1.00000000
Zscore(粉丝)	490	-.68013	5.16499	.0000000	1.00000000
Zscore(微博数)	490	-1.07811	3.93815	.0000000	1.00000000
Zscore(微博等级)	490	-4.02520	1.79293	.0000000	1.00000000
Zscore(爱慕值)	490	-.25532	11.61276	.0000000	1.00000000
Zscore(微博会员)	490	-4.96422	.52172	.0000000	1.00000000
有效个案数（成列）	490				

利用 SPSS 进行系统聚类。选取按 Z 得分标准化后的平方欧式距离作为距离衡量标准，采用组间链接法进行聚类，得到集中计划与谱系图。（由于样本较大，不在正文展示）。

1.2 确定聚类数

为确定分类数，我们利用集中计划中的聚合系数，研究其与分类数之间的关系

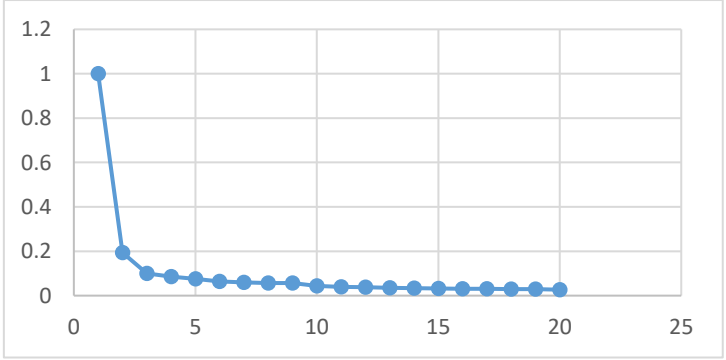

图 1 聚合系数随分类数的变化图

通过上图可以发现,分类数小于等于 3 时,聚合系数随分类数的增加下降迅速;分类数大于等于 4 后，聚合系数平缓地趋于零。由此，我们认为分类数应该选取 3 或者 4。分成 3 类的聚合系数为 0.10，分成 4 类的聚合系数为 0.08，从统计意义上研究，两者的差别不大。接下来我们结合现实意义，进一步确定我们的分组。

分成 3 类时，分组结果为

表 4 分组结果一

组别	成员
第一组	叶璇、蔡康永、金莎（共 3 人）
第二组	TFBOYS-易烊千玺、鹿晗（M 鹿 M）、TFBOYS-王俊凯（共 3 人）
第三组	其他（共 483 人）

分成 4 类时，分组结果为

表格 5 分组结果二

组别	成员
第一组	叶璇、蔡康永、金莎（共 3 人）
第二组	TFBOYS-易烊千玺、鹿晗（M 鹿 M）、TFBOYS-王俊凯（共 3 人）
第三组	文章（文章同學）、邓超、张杰、郭德纲、唐嫣、胡歌、王力宏、赵薇、林心如、杨幂、赵丽颖、何炅、Angelababy、谢娜、林志颖（夢想家林志穎）、陈坤、范冰冰、姚晨（共 18 人）
第四组	其他（具体名单见附录，共 466 人）

比较以上分组结果，可以看到，分成 3 类时大多数样本之间难以区分，反映的信息较少。分成 4 类可以对分成 3 类的结果进一步细分，且新生成的第三组特征明显，均是我们耳熟能详的名人。将第三组和第四组合并会造成大量信息的丧失。综上所述，我们最终确定分类数为 4，分类结果如表所示。

1.3 分析聚类结果

提取四组均值进行比较

表 6 分类结果均值比较

	关注	粉丝	微博数	微博等级	爱慕值	会员等级
第一类	1887.00	16848678.33	1468.67	36.67	16669.33	6.00
第二类	148.33	24510028.00	523.33	34.67	5428928.00	6.00
第三类	385.72	67229209.33	3994.11	38.67	302388.22	5.94
第四类	314.76	8296593.27	1992.13	32.81	80551.41	5.50

四组之间在微博等级和会员等级上相差不大，在其他指标上具有显著特征。

第一类“艺术家”（叶璇、蔡康永、金莎等 3 人）

他们关注数极高，约为其他组的 6 倍甚至 13 倍；他们的粉丝数排名第三，微博数排名第二，表明其在社会上的被关注度处于中偏上水平；但他们的爱慕值排名倒数第一，远远低于其他组别。可见，第一类属于在社会上有广泛知名度，但忠实粉丝数目较少的明星。他们出道时间较早，艺术作品较多，个人特色鲜明。

第二类 “流量王”（TFBOYS-易烊千玺、鹿晗（M 鹿 M）、TFBOYS-王俊凯等 3 人）

他们的关注数和微博数最少，表明他们在微博上的活跃度不及其他组，这可能与该组明星的偶像身份有关。他们的粉丝数排名第二，且与第一有着较明显的差距；但爱慕值却在所有组别中遥遥领先。总的来说，第二类有着不小的被关注量，粉丝群体中多为愿意为他们花钱刷榜的铁粉，属于当代流量“小鲜肉”类型。

第三类“大明星”（文章（文章同學）、邓超、张杰、郭德纲、唐嫣、胡歌、王力宏、赵薇、林心如、杨幂、赵丽颖、何炅、Angelababy、谢娜、林志颖（夢想家林志穎）、陈坤、范冰冰、姚晨等 18 人）

他们的粉丝数在所有类别中排名第一，且远远超过其他类别，是最受关注的一类明星。此外，他们的微博数也最多，超过第二名微博数的两倍，关注数位列第二。可见此类明星乐于在微博上分享自己的生活，与大众互动。但他们的爱慕值和收到花数仅位列第二，且与位列第一的第二类“小鲜肉”有较大的差距。流量是近些年兴起的衡量明星影响力的指标，而属于第三类的“大明星”们均出道超过五年，这与现实相符。

第四类“普通艺人”（周子扬等 466 人）

他们的关注数、微博数等处于一般水平；粉丝数最少，较少受到大众的关注。他们的爱慕值与粉丝数的比值相对较高，仅次于“流量王”，这也符合当前娱乐圈明星发展的趋势。

综上所述，我们通过系统聚类法，将样本分成“艺术家”、“流量王”、“大明星”、“普通艺人”这四类。这种分类符合现实情况，也能为我们研究明星发展提供见解。

2.快速聚类法

2.1 确定聚类数 k 值

快速聚类法中对类数目 k 值的确定依赖于实践经验的积累以及对系统聚类法聚类结果的参考，我们将分别从这两方面来讨论 k 值的选择。

首先，从经验出发，直观来看，我们的选择的变量（关注、粉丝、微博数、微博等级、爱慕值、微博会员）从艺人微博的活跃度及人气两方面描述了明星微博的基本情况，这两个方面则可以构成四种组合。

如下表所示：

表 7 明星微博类别

活跃度 / 人气	高活跃度	低活跃度
高人气	1	2
低人气	3	4

故从经验上，根据变量特征，我们倾向于将样本聚为四类。

其次，根据上述系统聚类方法下对聚合系数的研究以及与现实意义的结合，我们验证了将样本分为四类的直观经验是可靠的，是经得起检验的。

出于以上两方面考虑，我们决定在快速聚类法中，将 k 值设置为 4。

2.2 聚类结果展示与分析

2.2.1 初始聚类中心

表8 初始聚类中心

	1	2	3	4
Zscore(关注)	-.75699	1.39692	8.44976	-1.31571
Zscore(粉丝)	.67736	5.16499	-.10718	-.67252
Zscore(微博数)	-.92707	3.91541	-.61416	-.67534
Zscore(微博等级)	-.14644	1.10030	.13061	-4.02520
Zscore(爱慕值)	11.61276	-.05643	-.20630	-.25085
Zscore(微博会员)	.52172	.52172	.52172	-1.67265

由上表可以看出，这4个形心（“种子点”）之间的间距很大，几乎不可能在同一个类中，故聚类结果相对可靠。

2.2.2 样本分类情况

表 9 样本分类结果

组别	成员
第一组	TFBOYS-易烊千玺、鹿晗（M 鹿 M）、TFBOYS-王俊凯（共 3 人）
第二组	范冰冰、姚晨、贾乃亮、主播李湘、高圆圆、邓超、范范范瑋琪、萧敬騰-LION 獅子合唱團、GEM 鄧紫棋等（共 80 人）
第三组	林依轮、黄小蕾、阿雅、演员郭晓东、刘芸、胡定欣、Host 华少、钟丽缇 Christy、六小龄童等（共 111 人）
第四组	黄柏钧 Denny、芮伟航、黄梦莹 maggie、吴京、张静初、X 玖少年团-焉栩嘉 ziazia 等（共 296 人）

2.2.3 最终聚类中心

表10 最终聚类中心

	1	2	3	4
Zscore(关注)	-.71920	.34269	1.05055	-.47929
Zscore(粉丝)	.89370	1.69231	-.28066	-.36119
Zscore(微博数)	-.82836	1.17375	.55440	-.51673
Zscore(微博等级)	.22296	.70550	.55618	-.40150
Zscore(爱慕值)	11.19693	.06377	-.16824	-.06763
Zscore(微博会员)	.52172	.49429	.30426	-.25298

从上述“最终聚类中心”表格可以看出，四类微博用户样本在活跃度与人气两方面分别有自己的特征，与直观预期一致。

每一类的对应结果如下表所示：

表11 样本分类结果

活跃度 人气	高活跃度	低活跃度
高人气	2（大明星）	1（流量王）
低人气	3（艺术家）	4（普通艺人）

对各组分类结果的描述如下。

第一类“流量王”（TFBOYS-易烊千玺、鹿晗（M鹿M）、TFBOYS-王俊凯等3人）

关注数和微博数最少，微博等级较低，表明他们在微博上的活跃度较低，粉丝数较高，却远低于“大明星”，爱慕值遥遥领先，属于低活跃高人气型

第二类“大明星”（范冰冰、姚晨、贾乃亮、主播李湘、高圆圆、邓超、范范范玮琪、萧敬腾-LION獅子合唱團、GEM鄧紫棋等80人）

粉丝数在所有类别中排名第一，且远远超过其他类别，他们的微博数也最多，超过第二名微博数的两倍，关注数位列第二，但他们的爱慕值仅位列第二，且与位列第一的第二类“小鲜肉”有较大的差距，属于高活跃高人气型。

第三类“艺术家”（林依轮、黄小蕾、阿雅、演员郭晓东、刘芸、胡定欣、Host华少、钟丽缇Christy、六小龄童等等111人）

关注以及微博数较多，微博等级与微博会员也较为突出，说明他们在微博上的活跃度较高；但粉丝与爱慕值较低，属于高活跃低人气型。

第四类“普通艺人”（黄柏钧Denny、芮伟航、黄梦莹maggie、吴京、张静初、X玖少年团-焉栩嘉ziazia等296人）

他们的关注数、微博数等处于一般水平；粉丝数最少，较少受到大众的关注。属于低活跃低人气型。

2.2.4 方差分析

表 12 样本标准化后的方差分析

	聚类		误差			
	均方	自由度	均方	自由度	F	显著性
Zscore(关注)	67.150	3	.592	486	113.492	.000
Zscore(粉丝)	92.956	3	.432	486	214.993	.000
Zscore(微博数)	75.142	3	.542	486	138.553	.000
Zscore(微博等级)	40.673	3	.755	486	53.864	.000
Zscore(爱慕值)	126.978	3	.222	486	571.057	.000
Zscore(微博会员)	16.527	3	.904	486	18.280	.000

从方差分析表可以看出，所选的六个变量对分类贡献均显著，说明之前的变量选择过程较为科学，结果也较为可靠。

特别地，与系统聚类结果相比，快速聚类方法得到的结果虽然在每一类的具体人数上有较大差异，但每一类的特质以及分类思路都几乎相同，二者的结论大体相似，并且互相补充，也从侧面说明了两种聚类结果的可靠性。

五、 判别分析

在对明星微博进行聚类后，我们又从微博上搜集了几位近期新出现的热度较高的明星的相关数据，尝试通过判别分析对这几位明星所属类别进行判别。

（注：为消除量纲的影响，在此我们对于所有数据都进行了标准化处理。）

1. 确定先验概率

利用原数据进行一次判别分析，以验证判别的可行性并确定采用的先验概率。

1.1 以均等先验概率进行判别

表 13 均等先验概率判别结果

		案例的类别号	预测组成员				合计
			1	2	3	4	
初始	计数	1	3	0	0	0	3
		2	0	70	10	0	80
		3	0	0	104	7	111
		4	0	2	8	286	296
	%	1	100.0	0.0	0.0	0.0	100.0
		2	0.0	87.5	12.5	0.0	100.0
		3	0.0	0.0	93.7	6.3	100.0
		4	0.0	0.7	2.7	96.6	100.0

由结果可知，该判别对分组案例中的 94.5%进行了正确的分类，故我们认为从判别正确率角度而言该方法可行。

1.2 以个案所占比例为先验概率进行判别

表 14 非均等先验概率判别结果

		案例的类别号	预测组成员				合计
			1	2	3	4	
初始	计数	1	3	0	0	0	3
		2	0	63	11	6	80
		3	0	0	80	31	111
		4	0	0	0	296	296
	%	1	100.0	0.0	0.0	0.0	100.0
		2	0.0	78.8	13.8	7.5	100.0
		3	0.0	0.0	72.1	27.9	100.0
		4	0.0	0.0	0.0	100.0	100.0

该判别方式仅仅对分组案例中的 90.2%进行了正确分类，故我们选择均等的先验概率进行判别。

2. 判别组均值均等性检验

表 15 组均值均等性检验

	Wilks 的 Lambda	F	df1	df2	Sig.
关注	0.588	113.492	3	486	0.000
粉丝	0.430	214.993	3	486	0.000
微博数	0.539	138.553	3	486	0.000
微博等级	0.750	53.864	3	486	0.000
爱慕值	0.221	571.057	3	486	0.000
微博会员	0.899	18.280	3	486	0.000

可见所有指标各组中心都显著不相等，故我们认为该判别有意义。

3. 典型判别函数

$$y_1=0.014x_1-0.32x_2+0.091x_3+0.062x_4+3.922x_5-0.138x_6+0.182$$
$$y_2=0.282x_1+1.06x_2+0.879x_3-0.027x_4+0.121x_5+0.244x_6-0.002$$
$$y_3=1.047x_1-0.972x_2+0.244x_3+0.043x_4+0.288x_5+0.242x_6+0.005$$

其中，关注为x_1，粉丝为x_2，微博数为x_3，微博等级为x_4，爱慕值为x_5，微博会员为x_6

对以上三个判别函数的描述如下：

表 16 判别函数描述

特征值				
函数	特征值	方差的 %	累积 %	正则相关性
1	3.726[a]	55.5	55.5	0.888
2	2.216[a]	33.0	88.5	0.830
3	.770[a]	11.5	100.0	0.660
Wilks 的 Lambda				
函数检验	Wilks 的 Lambda	卡方	df	Sig.
1 到 3	0.037	1593.380	18	0.000
2 到 3	0.176	841.699	10	0.000

3	0.565	276.371	4	0.000

以上三个函数可提取全部的方差信息，且各组均值在三个函数上存在显著差异，我们保留以上三个判别函数，并利用以上三个判别函数对新变量进行判别

4. 对新样本进行判别

我们所新加入的数据如下（以下数据均已进行标准化）

表 17 新样本数据

	关注	粉丝	微博数	微博等级	爱慕值	微博会员
朱一龙	-0.44371	0.08409	-0.78227	0.96149	17.6459	1.61779
陈立农	-0.96705	-0.04947	-1.0376	-2.05946	4.86863	0.51929
华农兄弟	-1.21046	-0.62177	-1.00128	-2.74604	-0.19159	-0.57921
杨超越	-0.63033	-0.40304	-0.86087	-0.6863	0.81519	-0.57921
章子怡	0.07152	1.10609	-0.42719	0.54954	-0.18502	0.51929

利用判别函数所计算出的得分以及判别分组如下：

表 18 判别得分与分组

	第一判别函数得分	第二判别函数得分	第三判别函数得分	预测分组
朱一龙	69.12813	1.73945	4.79029	1
陈立农	18.98624	-0.47947	0.22809	1
华农兄弟	-0.57004	-1.96085	-1.21653	4
杨超越	3.45841	-1.37683	-0.40817	4
章子怡	-0.97268	0.89369	-1.00393	4

由以上结果，我们将朱一龙、陈立农分入组 1“流量王”（高人气低活跃），将华农兄弟、杨超越、章子怡分入组 4“普通艺人“（低人气低活跃）。

然而在我们的判别函数当中，包含了绝大多数方差的判别函数一和二中“爱慕值”一项所占比重极为突出，因此我们也可认为该结果相对其他因素而言，受到“爱慕值”一项的影响较大，因此一些知名度较高的明星比如章子怡并未能通过我们的分组以及判别体现出来，所以我们认为该判别模型对新样本的判别能力不是很强。

总结：通过多元统计方法并运用 SPSS 分析软件去探索我们所关注的娱乐圈是一件很有意义的工作，聚类与判别分析结果符合预期给我们带来巨大的喜悦。通过以上分析，我们可以看到各类明星在“流量称王”时代所处位置，然而真正有强大流量和吸金能力的明星还在少数，希望各位明星继续努力。