e企查

第十一届中国大学生服务外包创新创业大赛

基于加权多层K-Means的企业分类系统e企查

使用手册

团队成员：

鲍锋雄 程凯 陈振乾

陆纪慧 阳璐

e企查信息技术有限公司

大熊维尼队

指导教师：郑建炜

目录

一、前台使用手册

图 1-1：前台首页

如图 1-1 所示，用户输入网址后首先进入系统首页，系统首页分为三个部分

①用户注册登录以及管理员的后台登录部分

②搜索部分

③展示热搜企业部分

用户可以通过搜索企业全名、企业关键词和一些其他关键信息获取想要寻找的企业信息，搜索效果如图 1-2 所示：

图 1-2：企业关键词搜索

当用户输入企业关键词时（由于所给数据中的企业名进行了加密处理，所以演示中所使用的企业

名均为加密后的企业名字串），系统会根据用户的输入进行智能补全预测用户想要寻找的企业并在下方的推荐框中显示出来，若用户想要找的企业在推荐框中，则可以直接点击进入企业详情界面

图 1-3：验证码识别

点击搜索后，用户需要完成滑块拼图验证码，这是为了反爬虫，完成后跳转到企业搜索结果页面，如图 1-4 和 1-5 所示。

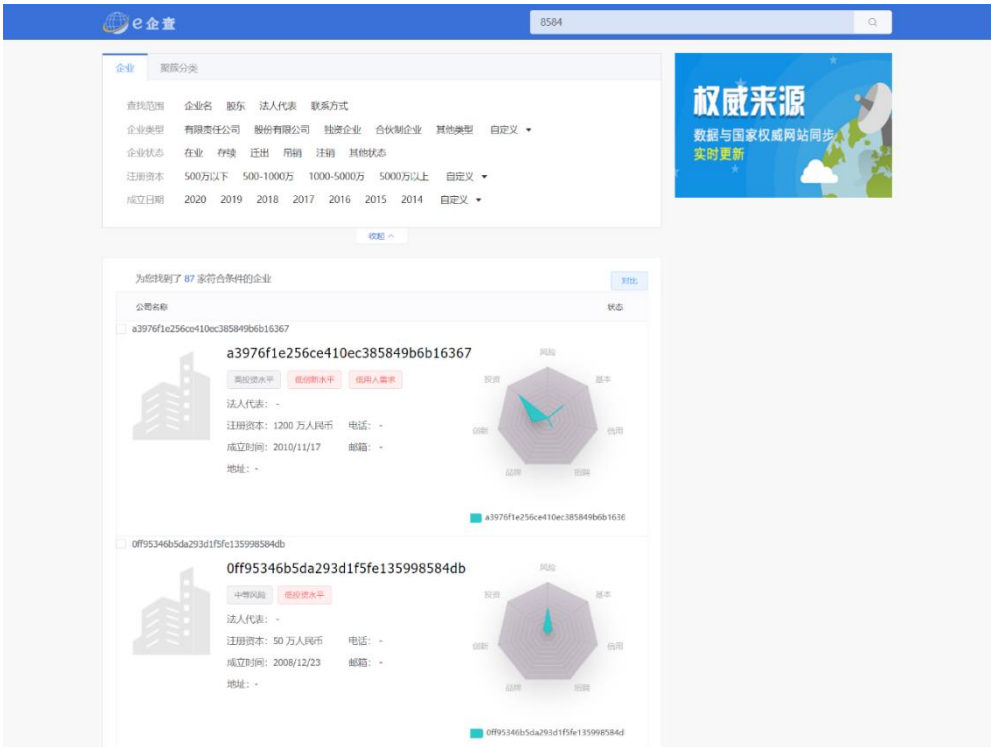


图 1-4：搜索结果页面 1

图 1-5 搜索结果页面 2

在企业搜索结果界面将列出含有搜索关键词企业的相关信息，包括企业照片、企业基本信息、以及通过聚类算法分类后根据簇特征而打上的标签，以及对企业评价的雷达图，该雷达图可以直观的表达对应企业各个指标的好坏，可以帮助用户更加全面的评估企业的价值。

图 1-6 ： 企业筛选选项

如图 1-6 所示，用户可以在关键词搜索结果之上进一步筛选结果，在图 1-6，是筛选标签。图 1-7 所示，是筛选的企业结果。

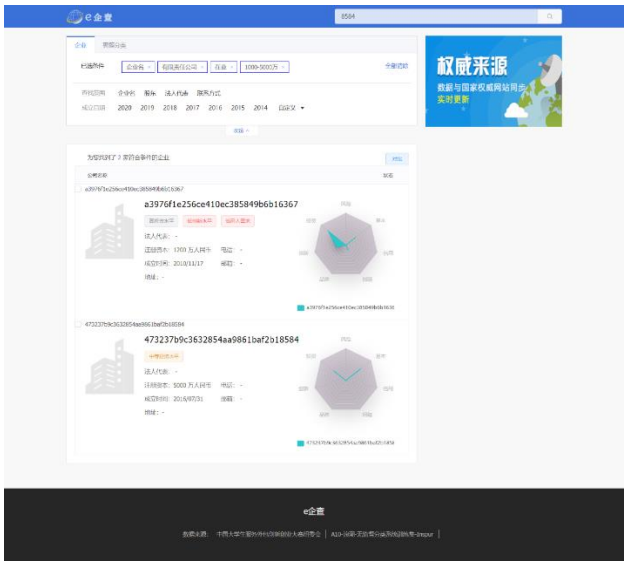


图 1-7：企业筛选结果

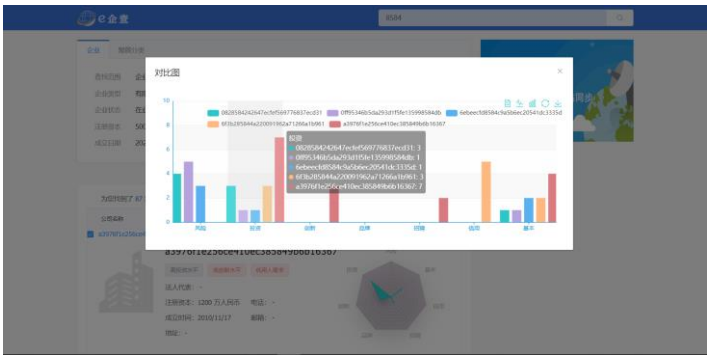


图 1-8：选择多个企业进行对比

用户可以在搜索结果界面上选择多个企业，对它们的评价属性进行对比，以柱状图的形式展现，最多可勾选 5 个企业。

点击聚簇分类标签页，用户可以根据企业各个模块的等级进行筛选。

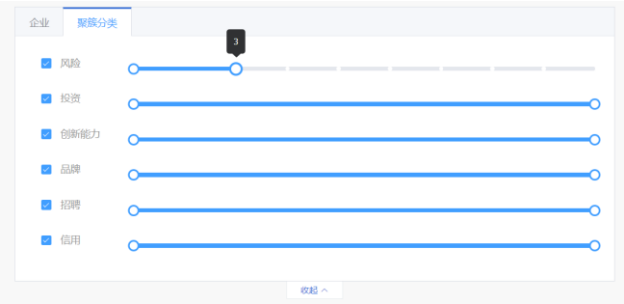


图 1-9：模块等级筛选

用户通过滑动滑块可以筛选对应属性的等级（等级为 0-10），勾选选择框表示筛选该属性，否则将排除该属性。如图 1-10 所示，是筛选结果。

图 1-10：企业筛选结果

用户点击企业条目可以进入企业详情页面

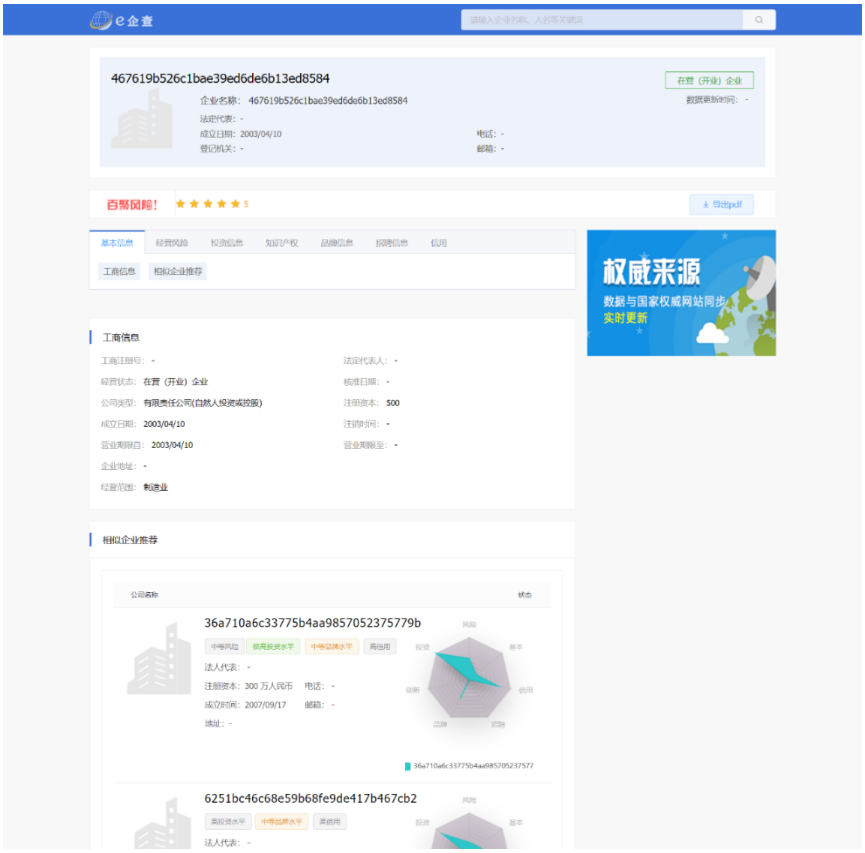


图 1-11：企业详情页面

在企业详情页面，首先是企业的一些最基本的信息，然后是企业的总分，以评星的形式展示，最直观的表达该企业好坏，在右边有一个导出企业报告 pdf 的按钮，接着用户可以依次查看企业的基本信息、企业经营风险、企业投资信息、知识产权信息、企业品牌信息、企业招聘信息以及企业信用。如图 1-12 所示。

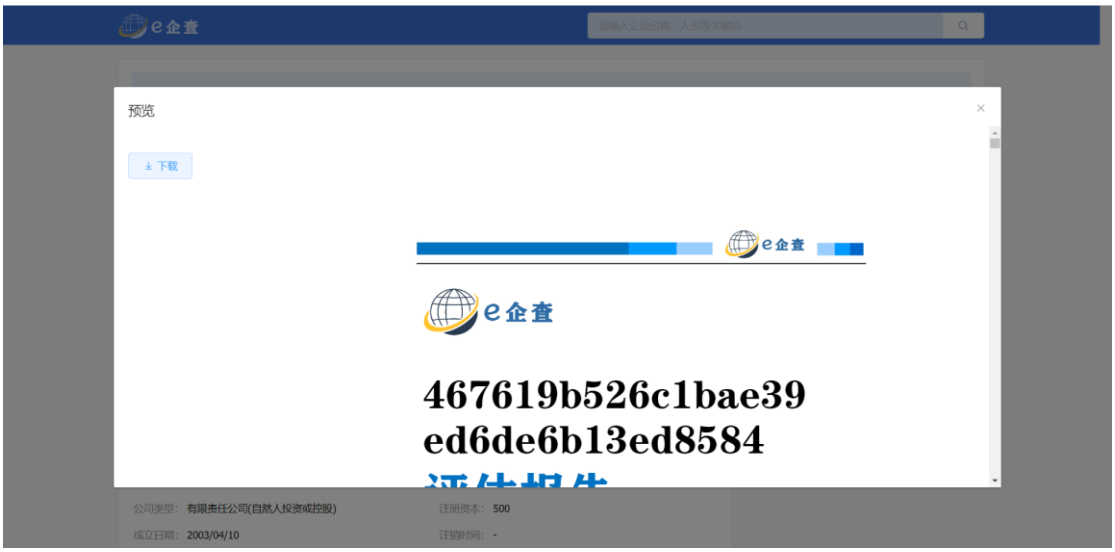


图 1-12：查看企业的评价报告

图 1-13：企业的评价报告导出为 pdf

用户点击导出 pdf 按钮后可以先预览企业报告，然后选择导出企业报告 pdf，如图 1-13 所示。

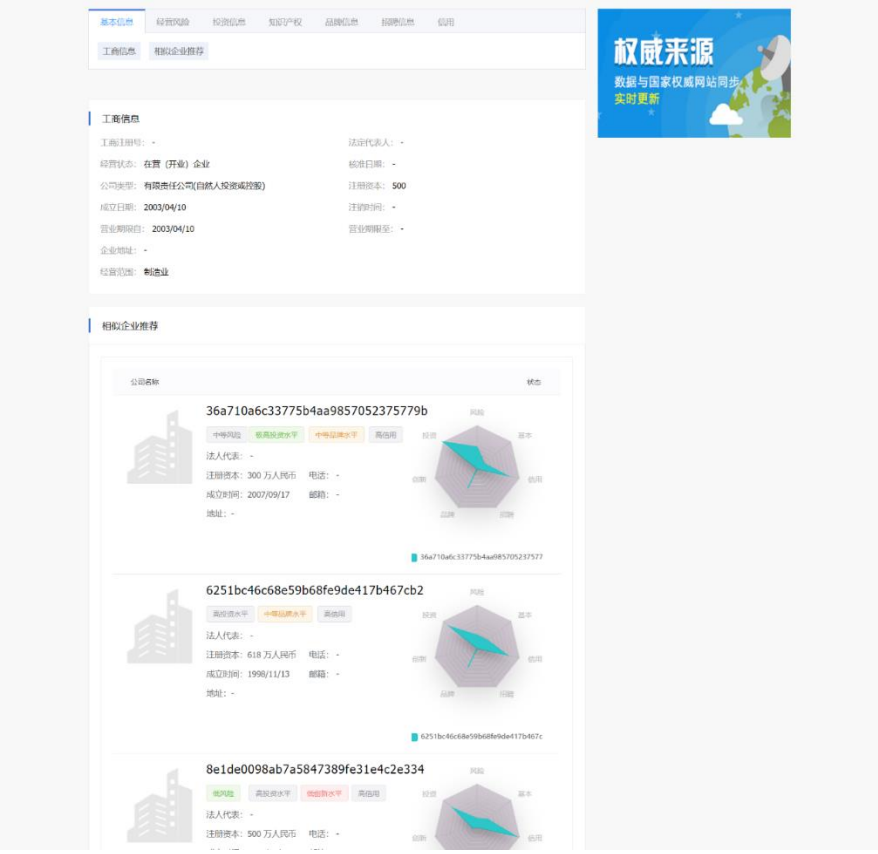


图 1-14：企业详情页面-基本信息

如图 1-14 所示，在企业基本信息标签页，用户可以查看企业的工商信息以及系统推荐的相似或相关企业，推荐企业以表的形式展现给用户，旁边的属性雷达图可以直观的展现推荐企业之间的联系。

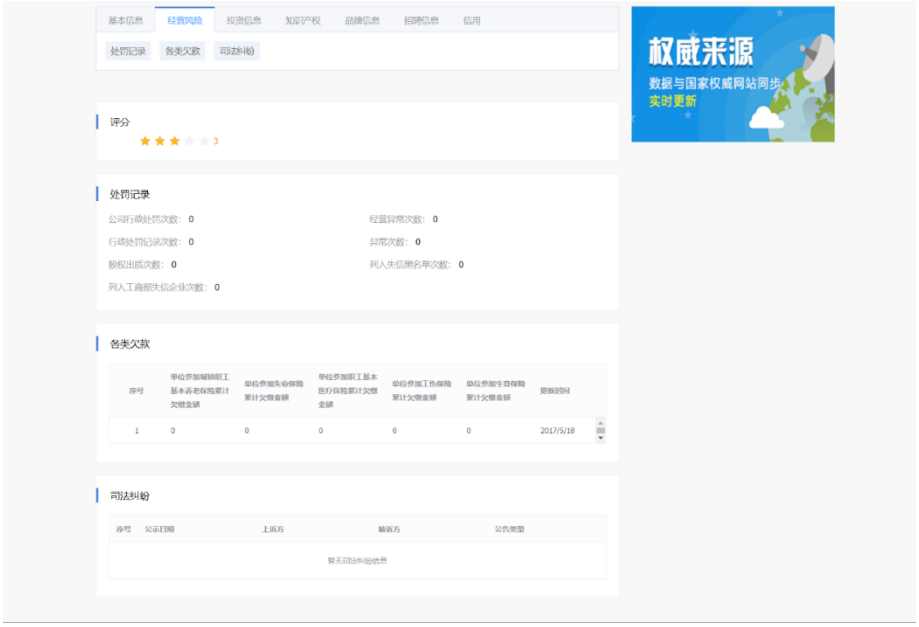


图 1-15：企业详情页面-经营风险

如图 1-15 所示，用户可以依次查看企业的风险评分（分数越高，风险越小）、企业受到的处罚记录、企业各类欠款的记录信息以及企业司法纠纷的记录信息。同理 1-16 至 1-20，可以查看企业知识产权、品牌、招聘、信用模块下的详细信息。

图 1-16：企业详情页面-投资信息

图 1-17：企业详情页面-知识产权

图 1-18：企业详情页面-品牌信息

图 1-19：企业详情页面-招聘信息

图 1-20：企业详情页面-信用

如图 1-21 至 1-23 所示，演示了如何通过上传 excel 或者 csv 文件，执行批量查询过程。

图 1-21：批量查询企业

图 1-22：下载 excel 示例文件

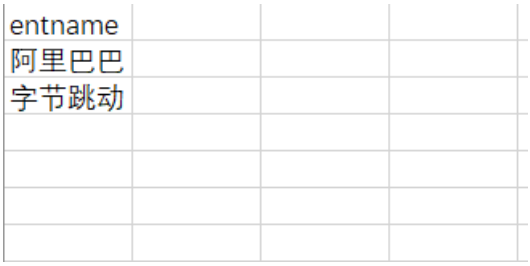


图 1-23：excel 示例

如图 1-24 所示，系统对于那些频繁被搜索的企业，提供了企业热搜功能。

图 1-24：首页热搜企业

如图 1-25 所示，用户点击后台登录按钮，页面将跳转到后台登录页面

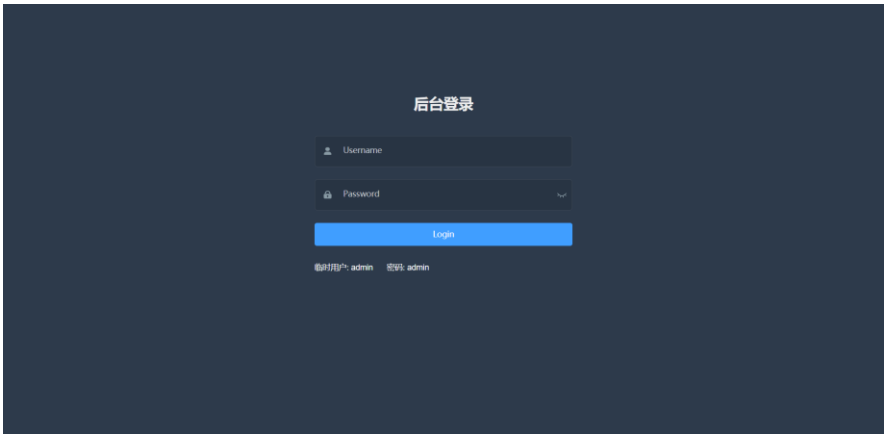


图 1-25：后台登录页面

二、后台使用手册

后台功能是帮助管理员或企业用户管理和调整模型参数权重、对聚簇结果和相关指标数据进行可视化展示、以及对输入测试样例进行预测分类并打上标签

2.1 控制台

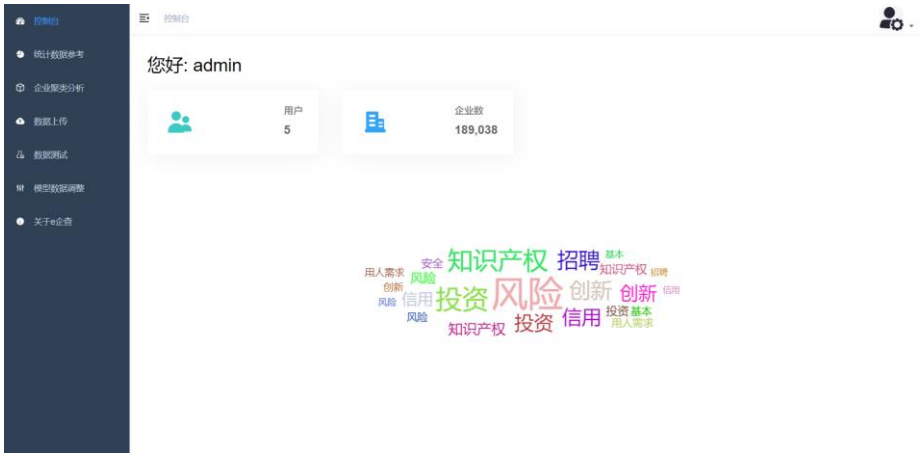


图 2-1 控制台

如图 2-1 所示，控制台显示当前系统的用户数，以及数据库中存储的企业个数。

2.2 统计数据参考

如图 2-2 所示，这是对聚类相关结果参数的图表化展示，可以直观的展现出在不同簇中的每个模块中的每个等级的占比情况。例如图 2-2 中，选中了 1 号簇，总分模块，可以显示总分模块不同等级在该簇内所占的百分比。

图 2-2：企业簇内模块分析

通过鼠标指针在曲线图上滑动，扇形图会根据曲线中所对应的属性而动态的改变，从而直观的展现出在该簇中，指定属性中的等级划分情况。

图 2-3：选择指定簇进行分析

如图 2-3 所示，通过左上方的选择框，用户可以选择查看指定簇中相关数据的分布情况。

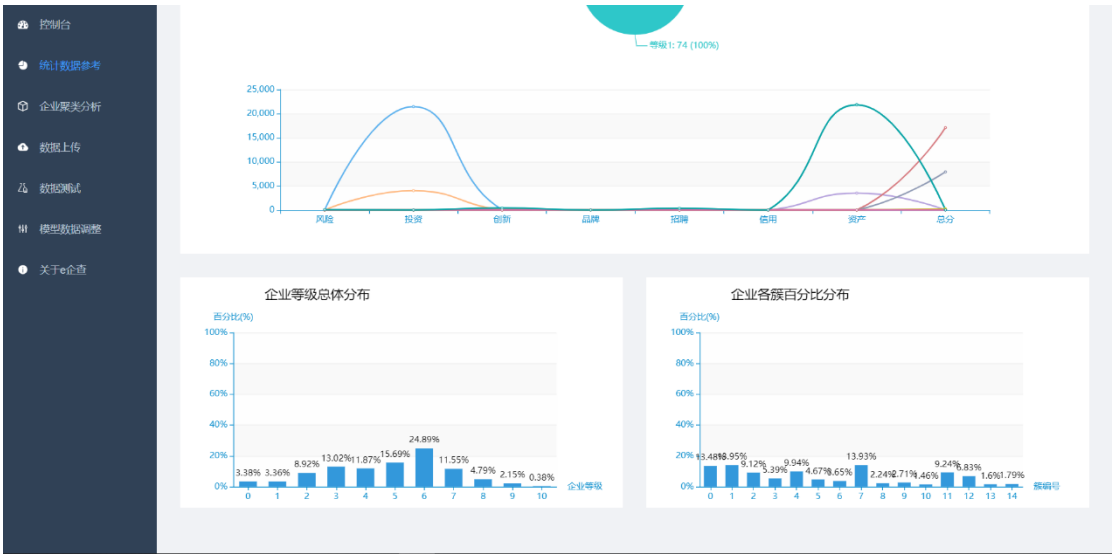


图 2-4：企业等级与企业各簇分布

如图 2-4 所示，企业等级最终划分为 0-10 个等级，左图显示的是每个等级在企业总体中的占比，总体呈现高斯分布。右图显示的是通过对 n 个企业，7 个模块属性，即 7×n 的矩阵进行加权聚类后，得到的聚类标签，共划分为 15 个簇（簇个数通过实验验证得到，为最优参数），显示每个簇所占百分比。

2.3 企业聚类分析

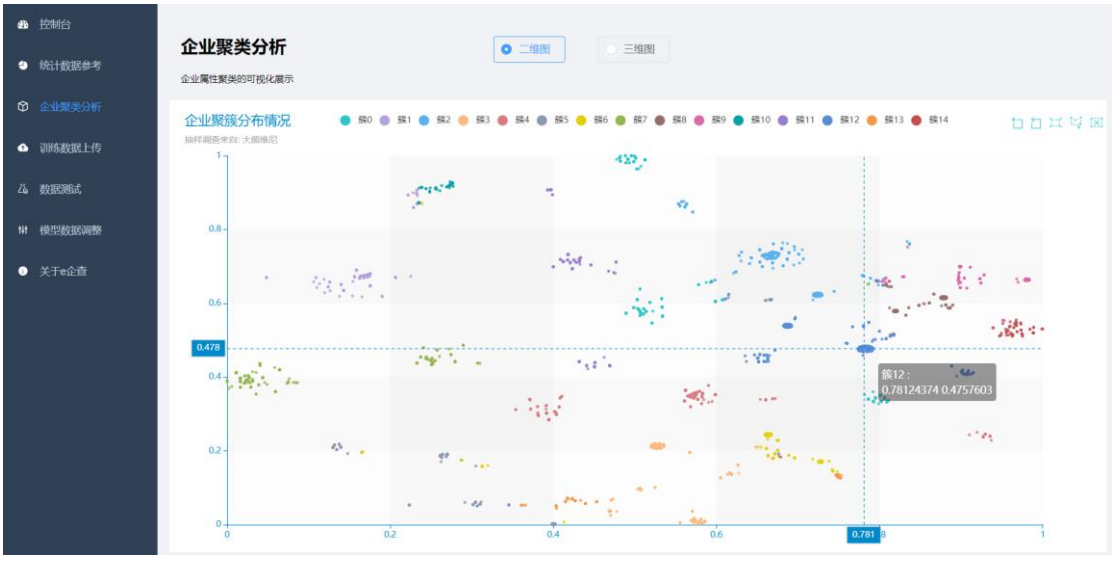


图 2-5：二维聚簇可视化

如图 2-5 所示，通过 PCA 降维将聚类得到的高维结果集降维到二维，将聚类结果以散点图的形式展现出来，散点图可以局部放大和区域选择。

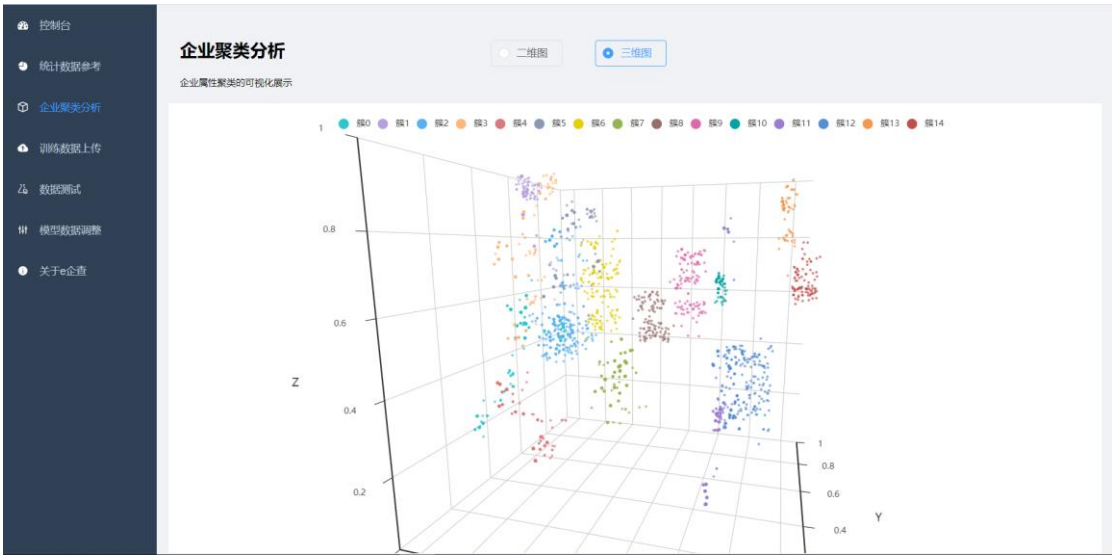


图 2-6：三维聚簇可视化

如图 2-6 所示，将聚类结果降到三维，通过三维立体图，更加直观的向用户展示聚类结果，三维图可以由用户任意拖动放大。

2.4 训练数据上传

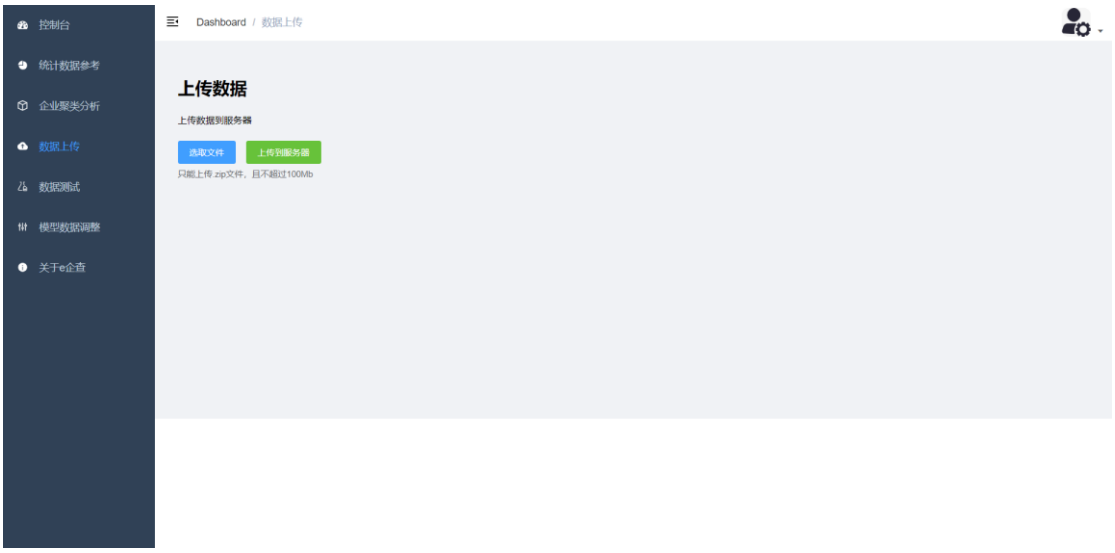


图 2-7：上传数据

如图 2-7 所示，点击选取文件，选择要上传的训练集，要求必须是 zip 压缩包的形式，文件格式和命名严格按照"服创大赛训练集-Inspur"中所给的数据集。点击上传到服务器按钮将文件上传到服务器，并存储到数据库中。

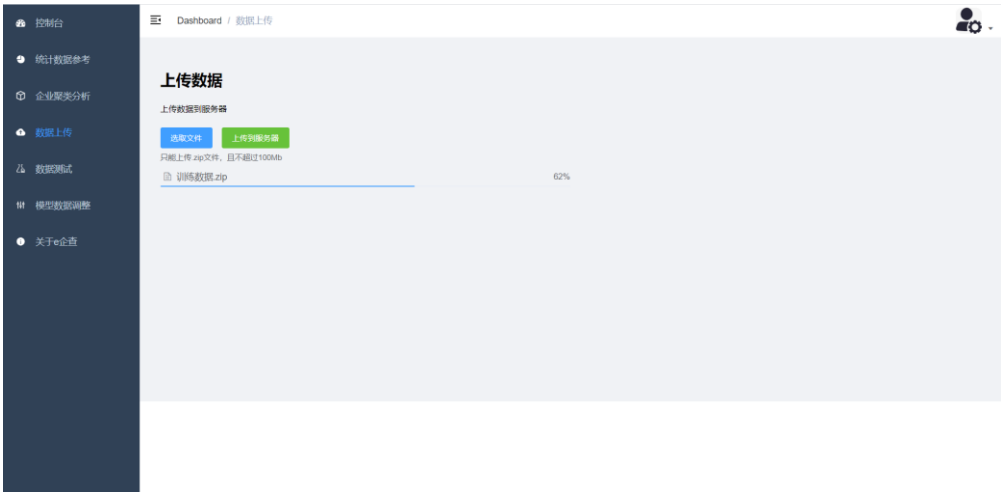


图 2-8：上传数据

如图 2-8 所示，上传数据会有进度条，显示上传进度。

2.5 数据测试

数据测试模块，提供了数据训练与测试功能。如图 2-9 所示，用户可以上传自己的训练集生成训练模型，并返回训练标签文件，和簇描述文件。预测有两种模式，一种是系统默认的模型，还有一种是通过训练测试生成的新模型进行预测。两种测试也是返回带标签训练文件和簇描述文件。

在训练或预测结束后，网页显示模型训练或预测时间，以及系统响应总时间。

图 2-9：数据测试页面

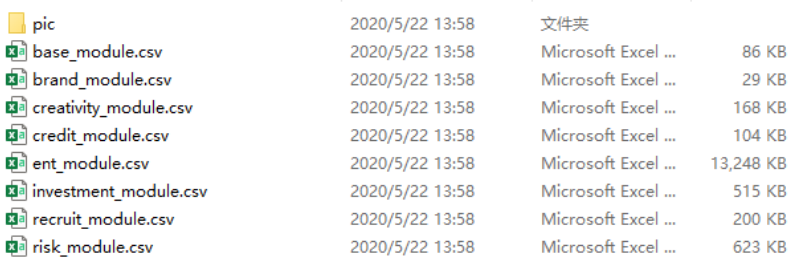
如图 2-10 所示，使用默认模型进行预测

图 2-10：使用默认模型进行预测

点击导入测试数据文件(要求为压缩包的形式)，当完成导入时下方会显示文件夹名，点击开始预测对测试数据进行预测， 请耐心等待几秒钟即可完成预测，并会有预测成功和下载结果文件的提示。

图 2-11：使用用户训练模型进行训练

如图 2-11 所示，需要训练数据时，请点击导入训练数据并选择训练数据文件(要求为压缩包的形式)导入，当完成导入时下方会显示文件名，点击上传进行训练请耐心等待几秒钟即可完成训练，并有训练成功的提示和相关用时。



pic	2020/5/22 13:58	文件夹	
base_module.csv	2020/5/22 13:58	Microsoft Excel ...	86 KB
brand_module.csv	2020/5/22 13:58	Microsoft Excel ...	29 KB
creativity_module.csv	2020/5/22 13:58	Microsoft Excel ...	168 KB
credit_module.csv	2020/5/22 13:58	Microsoft Excel ...	104 KB
ent_module.csv	2020/5/22 13:58	Microsoft Excel ...	13,248 KB
investment_module.csv	2020/5/22 13:58	Microsoft Excel ...	515 KB
recruit_module.csv	2020/5/22 13:58	Microsoft Excel ...	200 KB
risk_module.csv	2020/5/22 13:58	Microsoft Excel ...	623 KB

图 2-12：训练或预测下载结果

如图 2-12 所示，下载结果包含 7 个模块的 csv，以及最后一个总评的 ent_module.csv。以及簇描述文件 pic。

对其中的一个模块进行分析，以 risk_module.csv 为例，如图 2-13 所示。

entname	enforce_ar	is_punish_t	taxunpaidr	is_brap_typ	appellant_	defendant	is_justice_c	judge_new	is_except_t	pledgenun	is_justice_c	is_bra_type	risk_modul	risk_module_type
08f3eb544	1	0	0	0	3	0	0	0	0	0	0	0	2.7	4
0e940bcbc	2	0	0	0	0	0	0	3	0	0	0	0	5.1	6
150ac0eb3	1	0	0	0	0	0	4	0	0	0	0	0	5.8	7
27f75dff67	1	0	0	0	0	0	0	2	0	0	0	0	2.8	4
2feeb2d7c	4	0	0	0	0	0	0	0	0	0	0	0	7.2	8
31ebae895	1	0	0	0	0	0	0	0	0	0	0	0	1.8	3
3eb34efb7	1	0	0	0	0	0	1	1	0	0	0	0	3.3	4
40c9c7812	2	0	0	0	0	0	0	0	1	0	0	2	6.6	7
4372222d8	1	0	0	0	0	0	2	0	0	0	0	0	3.8	5
437dd86fe	1	0	0	0	0	0	1	0	0	0	0	2	4.8	6
497adc0c7	2	0	0	0	0	2	1	0	0	0	0	0	5.8	7
49dfd7751	1	0	0	0	0	0	0	2	0	0	0	0	2.8	4
4c7db6b8(	1	0	0	0	0	0	2	0	0	0	0	4	7.8	8
4cca907e0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4eb466db(	1	0	0	0	0	2	0	0	0	0	0	0	3	4
4f7cdab19	5	0	0	0	0	0	0	0	0	0	0	0	9	9

图 2-13：风险模块统计

risk_module 前的若干列，是风险模块作为建模指标的等级化结果，例如以第一个企业（08f3eb544996cdd517b1a7bd1d0e9bb0）为例，它的被执行金额等级为 1，原告次数等级为 3，风险模块加权总分为 2.7，风险等级为 4。其他 6 个模块的分析同理。

以模块总评 ent_module.csv 为例，是对企业 7 个模块进行总评，并给出企业加权总分，和企业综合等级，以及企业簇标签。

	A	B	C	D	E	F	G	H	I	J	K
	entname	risk_modul	investment	creativity_r	brand_mo	recruit_mo	credit_moc	base_mod	ent_modul	ent_modul	ent_inner_ty
	08f3eb544	4	0	0	0	0	0	0	-4.44	1	11
	0e940bcbc	6	0	0	0	0	0	0	-6.66	1	12
	150ac0eb3	7	2	0	0	0	0	0	-5.77	1	12
	27f75dff67	4	3	0	0	0	0	0	-1.44	2	3

图 2-14：模块总评统计

如图 2-14 所示，还是以 08f3eb544996cdd517b1a7bd1d0e9bb0 为例，它的风险等级为 4，其他个模块由于信息为空，所以都为 0，最后企业加权总分为-4.44（风险对企业总评起副作用，因此权重为负），得到最后企业等级为 1，归入到 11 号簇中。

由于 11 号簇这个概念较为模糊，我们提供了描述来描述 11 号簇，如图 2-15 所示，是簇描述文件：

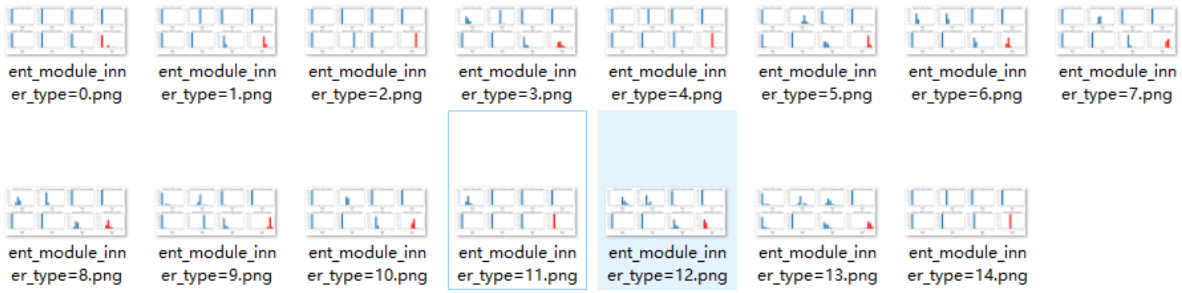


图 2-15： pic 目录下的簇描述文件

名称下的 ent_module_inner_type=x，表示 x 号簇的描述图片。以 11 号簇为例，如图 2-16 所示：

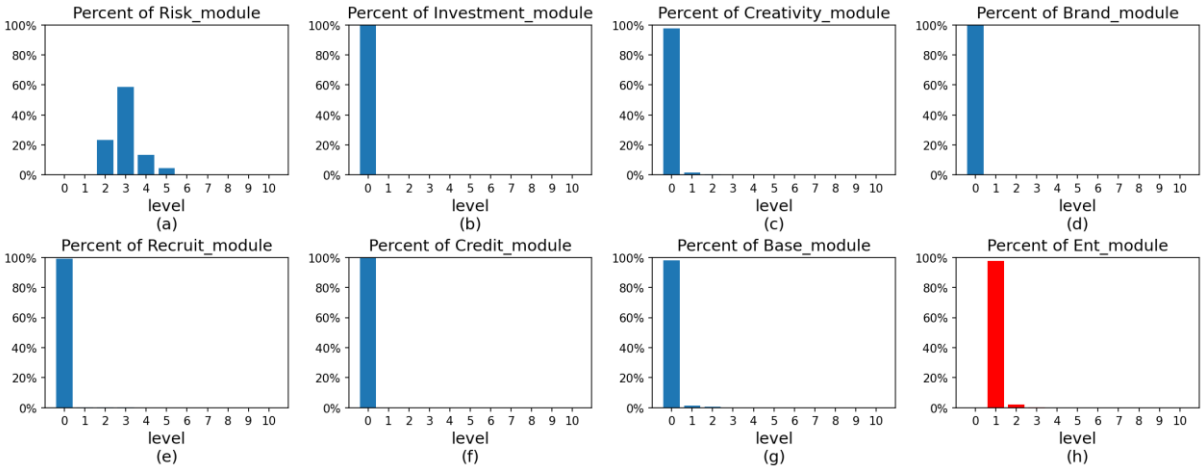


图 2-16：11 号簇描述

该描述文件表达的意思是，11 号簇中，风险、投资、创新、品牌、招聘、信用、基本信息 7 个模块，对应模块内每个等级的占比分布。例如 2-16 中图（a）表示 11 号簇中，风险等级基本在 2-5 的范围，以 3 号簇为主，其他模块基本等级为 0。最后的企业总评如图（h）所示，可以看出 11 号簇企业等级基本为 1。可以看出图 2-16 能概括 11 号簇的总体特征，08f3eb544996cdd517b1a7bd1d0e9bb0 这个企业也确实符合该簇的描述。

2.6 模型数据调整

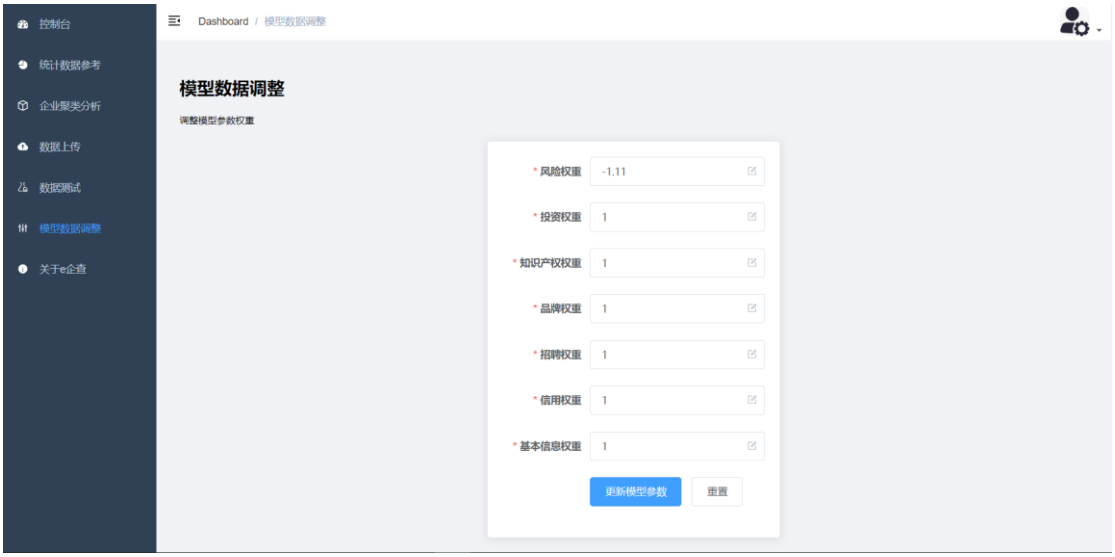


图 2-17：模型参数权重调整

如图 2-17 所示，管理员可以自定义模型的权重，点击更新模型参数可以修改模型参数的权重，在下次训练模型时生效。