打造智慧企业分类可视分析！

——小微企业智慧分类可视分析系统

队名：蓝色方

目录

1.项目介绍

1.1.项目摘要

金融科技是未来最核心的竞争力。在数字经济时代，科技赋能银行业务，将深刻改变银行业态，变革行业格局。大力投资金融科技，具备强大的金融科技实力的银行，才是面向未来的银行。金融科技的应用克服了传统金融信贷场景审核流程长、放贷慢的弊端，应用机器学习方法自动评估小微企业信用水平、企业还款能力等是金融科技在金融场景中的重要应用之一，然而其中存在着一些问题：

一是数据量大，企业这一信贷主体的数据覆盖互联网、政府、线上应用等来源的方方面面，来源广泛，在分析企业还款能力、信用水平过程中面临巨大的挑战。

二是企业数据维度丰富，高维数据直观表示难，当维数越来越多时，数据计算量急剧上升，所需的空间样本数会随着位数的增加而呈指数增长，分析和处理多维数据的复杂度和成本也呈指数级增长。

三是采用这样数据驱动的方法直接得到的分类结果缺乏可解释性，不容易获得用户的信任。

结合以上三大问题，我们在使用机器方法实现无监督分类的基础之上加入了可视化方法和可视交互技术，完成了我们的小微企业智慧分类可视分析系统。它是一个自动进行企业分类、构建企业分类画像，通过可视化与用户交互于一体的企业分类系统。我们知道传统的金融信贷场景需要领域专家去对企业数据进行分析考察，存在着人工成本过高，效率低下、缺乏公平性等问题，也比较容易出现人为失误。这样低效的信息获取速率不再适应互联网时代下信息的高速增长。我们的智慧分类系统首先对数据进行预处理，再使用无监督分类方法进行聚类，然后在 Web 端进行分类结果的可视化，用户还可以在 Web 端对公司进行多个或单个的类别信息查询。

针对问题一：我们合理分析有效企业数据及数据特征，针对这些数据进行更进一步的聚类和可视化工作，以高效、简介、支持交互的方式呈现高维、多层次企业数据。

针对问题二：我们采用对数据降维的方法，使用流形学习的方法将高维的数据呈现在二维平面中，实现了高维数据的直观展示。

针对问题三：我们使用可视化技术展示了更加底层的分类信息，交互技术使用户能够更好地探索数据之间的关联性，这在一定程度上让用户了解了无监督分类的机制和规律，增强了无监督模型的可解释性。

可视化方法和可视交互技术使用户和无监督的分类方法之间形成了友好的的信息交互，使用户不仅可以看到机器学习的分类结果，还增加了系统决策的可信度，给用户带来更多直观并且深刻的体验，适应在金融科技场景下人们对于机器学习模型可靠性和信任度的需求。

1.2. 需求分析

1.2.1. 问题分析

对企业多源数据、多维度进行深入挖掘，为企业构建企业画像、建立企业信用评分体系打下基础，用户可从企业的企业背景、经营能力、经营风险、发展状况等层面对企业进行了解。

1.2.2. 功能需求分析

我们将系统需求分为机器学习的聚类需求和可视化 Web 端需求。

- 聚类需求：

（1）针对无标识的企业数据进行数据预处理，特征筛选，特征提取等形成有效的训练样例及特征；

（2）针对提取的有效特征选择合适的无监督分类方法对小微企业数据进行分类，进行模型训练，模型要求实现小微企业群体的有效划分；

- 可视化需求：

（1）针对小微企业划分后各簇提取显著标签进行该簇的描述，要求标签合理且有效；

（2）企业无监督分类要求最终以完整系统的形式接收企业信息输入，展示企业划分簇类别、该企业所在簇的有效标签。

表 1-1 聚类功能一览表

	功能项	功能项描述
聚类	特征筛选	模型中使用的特征名由表格名与字段名拼接而来
	特征分类	若干个特征组合起来表征企业某一方面的特征
	K-Means 聚类	确定 k 值，选择质心之后聚类
	评价指标	用 Inertia、CH 指标和轮廓系数评价聚类结果好坏
	聚类结果分析	雷达图、折线图、平行坐标图来可视化聚

		类结果

表 1-2 Web 前端可视化

	功能项	功能项描述
Web 端	搜索框	支持输入单个企业和多个企业，与数据库相连。
	散点图	在散点图上加入交互以及鱼眼可视化
	基于力导向布局的词云图	在原始标签词云图基础上加上力导向布局
	雷达图	雷达图对聚类结果进行总结性分析
	信息展示框	与数据库相连，展示单个或多个企业具体信息

1.2.3. 非功能性需求

无监督分类可视系统的开发符合一般应用开发的基本原则规范，具体要求如下：

① 易用性

无监督分类可视系统的使用对象主要为商业银行及其员工，因而该系统应该具备简单易操作的特性。系统的界面设计应该简洁美观，符合大众的审美；系统的功能设计应符合一般逻辑，操作简单，易于使用。

② 易懂性

各个功能所显示的页面内容应充分遵从易懂性的原则，使用者可直观理解其内容，不可复杂难懂，否则很多用户容易放弃使用我们的软件。

③ 可变性

现在的银行中小企业贷款市场非常广阔，无监督分类可视系统需要不断紧跟时代的潮流，不断进行推陈出新。所以为了长远考虑，本系统的设计必须灵活可扩展，以便以后能在用户的需求发生变化的时候及时作出调整，满足用户新的需求。

④ 响应及时性

由于本系统需要对大量的数据进行处理、统计和分析，要求能够及时响应用户的操作指示，不可出现长时间无响应的状态，以提高用户对本系统的满意度。

1.2.4. 业务流程分析

(1) 业务流程

接包后，本团队先对现有企业数据无监督系统进行分析，结合可视化的优点，参考该项目所需要的基本功能以及要求综合考虑，设计出具体方案。

(2) 机器学习的功能：

- **特征筛选和分类：**原始数据集中特征过多，其中，很多特征或是太过稀疏，或是重要性不够，在进行聚类前应当予以删除或相应处理。组合若干特征起来表征企业某一方面的特征。
- **聚类：**通过机器学习的 K-Means 等方法得到企业数据聚类结果。
- **结果分析：**用 Inertia、CH 指标和轮廓系数评价聚类结果好坏，雷达图、折线图、平行坐标图来可视化聚类结果

(3) Web 前端的功能：

- **用户：**接收单个或多个企业信息输入，展示企业具体信息、划分簇类别、该企业所在簇的有效标签。

1.3. 项目介绍

根据项目所需实现的目标，我们主要设计了如图 1-1 所示的系统功能模块图：

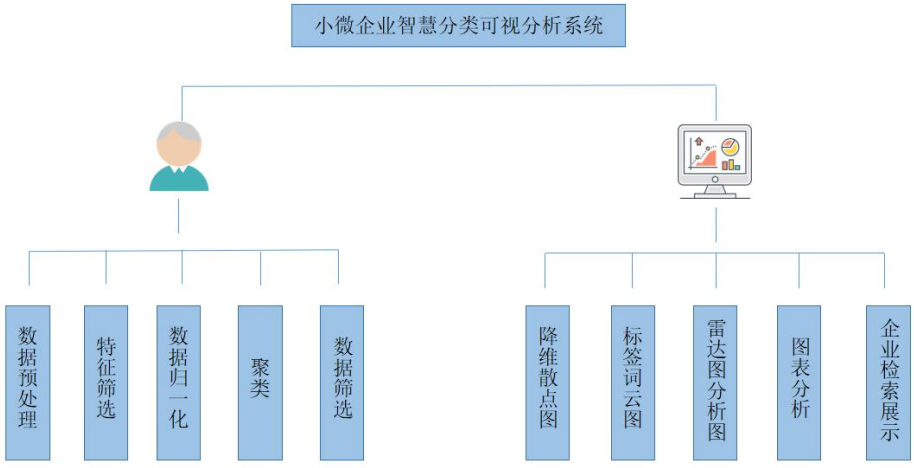

图 1-1 系统功能架构图

本团队设计了一套小微企业智慧分类可视分析系统设计方案。方案为应用中涉及的用户，根据企业需求，在前期提供核心的数据预处理、特征筛选、数据归一化聚类以及数据筛选工作。Web 端主要提供企业高维数据降维可视化、标签词云展示、可视图表多维分析企业信贷能力以及企业检索和类别、基本信息展示功能。

1.4.创新点分析

1.4.1.特征分类与汇总

原始的数据特征繁多，在数据处理阶段发现对结果影响因子较小的特征可以忽略。采用的特征中，我们发现若干个特征可以组合起来表征企业某一方面的特征。本项目中，我们为企业设计了五个方面的考量因素，分别是企业背景、企业经营能力、企业经营风险、企业司法风险、企业发展状况。

1.4.2.增强决策可信度

无监督学习可以让我们松开系在机器上的皮带，让机器嵌入数据，自主地进行发现和体验，寻找模式和联系并得出结论。但是无监督学习会有可解释性差的问题，在本项目中，我们通过降维可视化展示形成符合人的认知的形式，展现了更加底层的分类信息的同时增强机器学习的决策可信度。

1.4.3.多视图协调关联可视分析

可视化是人类认识、分析复杂数据的重要途径。企业数据具有高维、复杂、大众熟知率低的特点，借助可视化手段，实现多视图协调关联可视分析，将显示区域划分为多个试图，分别实现降维可视化、词云标签可视化、分析图表展示等模块的关联分析，从类别、企业多维度特征分析小微企业信贷能力。

1.5.产品展示

1.5.1.数据处理

(1) 数据预处理

原始的数据，特别是企业方面的数据，往往都会包含许多种类，并且存在着一些“坏”的数据，需要在进行正式地聚类前进行适当的预处理。本项目中的预处理包括以下几个操作：

- 特征处理

企业提供的数据中，许多特征对最后的聚类其实是没有帮助的，甚至会增大模型的复杂性，起到反作用。这些需要删除的特征大体上有两类，一类是数据太过稀少，在总样本数达到十几万的情况下，很多特征只有几百条、甚至几十条记录，而且它们的重要性也不是很高，因此需要删除；另一类虽然记录数很多，但本身无太大意义，如日期类的特征，这种也应删除。具体删除的特征可以参考表 1-3。

表 1-3 删除特征表

表格名称	字段	字段含义	备注
COMPANY_BASEINFO	EMPNUM	从业人数	参考意义不大
	ESTDATE	成立日期	参考意义不大
	CANDATE	注销时间	参考意义不大
	REVDATE	吊销时间	参考意义不大
	ENTSTATUS	企业状态	参考意义不大（因为最终只保留了企业状态为“在营企业”的企业）
	OPTO	经营(驻在)期限至	参考意义不大
	ENTTYPE	企业(机构)类型	参考意义不大
	ENTCAT	企业类别	参考意义不大
	INDUSTRYPHY	行业门类	参考意义不大
	REGCAPCUR	注册资本(金)币种	参考意义不大
	OPFROM	经营(驻在)期限自	参考意义不大
CHANGE_INFO	REMARK	备注	我们认为这张表提供的信息没太大参考意义，所以整张表的特征都予以剔除
	DATAFLAG	数据来源标志：1 核	

		准通过2删除或者驳回或者不予受理	
	ALTTIME	变更次数	
	ALTITEM	变更事项	
	ENTNAME	公司名称（加密后）	
	CXSTATUS	撤销状态：1：变更2:撤销变更3:已撤销变更	
	ALTDATE	变更日期	
	OPENO	业务编号	
ENTERPRISE_CONTRIBUTION	ENTNAME	企业名称	表中数据过于稀疏且无太大参考意义
	INVTYPE	投资人类型（股东类型）（19种）	
	CONFORM	出资方式(认缴)（货币、美元）	
	SUBCONAM	认缴出资额（认缴额万元）	
	CONPROP	持股比例	
	CONDATE	出资日期(认缴)	
ENTERPRISE_CONTRIBUTION_YEAR	SUBCONCURRENCY	认缴币种	表中数据过于稀疏且无太大参考意义

	ACCONDATE	实缴出资时间	
	SUBCONFORM	认缴出资方式	
	ANCHETYPE	行业分类	
	SUBCONDATE	认缴出资时间	
	ACCONCURRENCY	实缴币种	
	ACCONFORM	实缴出资方式	
	ENTNAME	企业名称	
	LIACCONAM	累计实缴额	
	LISUBCONAM	累计认缴额	
ENTERPRISE_GUARANTEE	PRICLASECKIND	主债权种类	表中数据过于稀疏且无太大参考意义
	PEFPERFROM	履行债务的期限自	
	IFTOPUB	是否公示此担保信息 1 是 2 否	
	PRICLASECAM	主债权数额	
	PEFPERTO	履行债务的期限至	
	GUARANPERIOD	保证的期间1期限2未约定	
	ENTNAME	企业名称	

	GATYPE	保证的方式 1 一般保证 2 连带保证 3 未约定	
	RAGE	保证担保的范围	
RECRUIT_QCWY	ENTNAME	企业名称	表中数据过于稀疏且无太大参考意义
	QCWYNUM	招聘记录条数	
RECRUIT_ZHYC	ENTNAME	企业名称	表中数据过于稀疏且无太大参考意义
	ZHYCNUM	招聘记录条数	
RECRUIT_ZLZP	ENTNAME	企业名称	表中数据过于稀疏且无太大参考意义
	ZLZPNUM	招聘记录条数	
ENTERPRISE_INSURANCE	CBRQ	参保日期	剔除了表中原有特征，但根据表中信息提炼了 1 个特征，详见 1.2 节内容
	XZBZ	险种标志	
	SBJGBH	社会保险经办机构	
	XZBZMC	险种标志名称	
	CBZT	参保状态	

	CBZTMC	参保状态名称	
	DWBH	单位编号	
	ENTNAME	查询企业	
ENTERPRISE_SOCIAL_SECURITY	ENTNAME	企业名称	因为欠缴金额大部分是 0，故认为参考意义不大
	UNPAIDSOCIALINS_SO210	单位参加失业保险累计欠缴金额	
	UNPAIDSOCIALINS_SO310	单位参加职工基本医疗保险累计欠缴金额	
	UNPAIDSOCIALINS_SO410	单位参加工伤保险累计欠缴金额	
	UNPAIDSOCIALINS_SO110	单位参加城镇职工基本养老保险累计欠缴金额	
	UNPAIDSOCIALINS_SO510	单位参加生育保险累计欠缴金额	
	UPDATETIME	更新时间	
PRODUCT_CHECKINFO_CONNECT	ENTNAME	企业名称	表中数据过于稀疏且无太大参考意义
	parent	企业产品被抽查的	

		合格率（未被抽查值为0，被抽查则值为0-1之间小数值）	
JUSTICE_DECLARE	ENTNAME	企业名称	剔除了表中原有特征，但根据表中信息提炼了2个特征，详见1.2节内容
	DECLAREDATE	公告时间	
	APPELLANT	上诉方(企业如果为上诉方，值为1，否则值为0）	
	DEFENDANT	被诉方（企业如果为被诉方，值为1，否则值为0）	
	DECLARESTYLE	公告类型	
JUSTICE_ENFORCED	ENTNAME	企业名称	
	RECORD_DATE	立案日期	参考意义不大
	CASE_NO	案号	参考意义不大
JUSTICE_JUDGE_NEW	ENTNAME	企业名称	剔除了表中原有特征，但根据表中信息提炼了1个特征，详见1.2节内容
	TIME	时间	

	TITLE	标题	
	CASETYPE	案件类型	
	JUDGERESULT	判决结果	
	CASECAUSE	案由	
	EVIDENCE	依据	
	COURTRANK	法院等级	
	DATATYPE	案由编码类型	
	LATYPES	司法类型	
JN_CREDIT_INFO	ENTNAME	企业名称	
	credit_grade	信用等级N+、B-、A、C、N、A-	参考意义不大
JN_TECH_CENTER	ENTNAME	企业名称	
	LEVEL_RANK	级别(省级2、市级1、企业名称不出现在该表则值为0)	参考意义不大
JN_SPECIAL_NEW_INFO	ENTNAME	企业名称	
	is_jnsn	是否是济南市专精特新中小企业（缺失值99.99%）	参考意义不大
INTANGIBLE_BRAND	ENTNAME	企业名称	
	ibrand_num	知识产权--商标申请次数	参考意义不大

WEB_RECORD_INFO	ENTNAME	企业名称	
	idom_num	企业是否拥有域名的知识产权	参考意义不大

保留的特征就是模型中使用的特征名，由表格名与字段名拼接而来，格式为<表格名|字段名>。

表 1-4 保留特征表

特征	说明	备注
company_baseinfo.csv|regcap	注册资本	对原始数据进行了处理，统一了单位
business_risk_abnormal.csv|is_bra	企业被列入经营异常的次数	
business_risk_all_punish.csv|is_brap	企业被行政处罚的次数（综合）	
business_risk_rightpledge.csv|pledgenum	出质股权次数	
business_risk_taxunpaid.csv|taxunpaidnum	企业累计欠税额	
administrative_punishment.csv|is_punish	企业被行政处罚的次数	
enterprise_insurance.csv|cbzt	企业给员工上的保险数目	cbzt与原先的同名字段含义不同，这是经过提炼后的特征，表示企业为员工买的保险数目。比如，五险一金要是都有的话，那就是6
enterprise_keep_contract.csv|is_kcont	被列为守合同重信用企业的次数	
ent_bid.csv|bidnum	中标次数	
ent_branch.csv|branchnum	分支机构数	
ent_investment.csv|investnum	对外投资次数	
ent_onlineshop.csv|shopnum	网店个数	
exception_list.csv|is_except	被列入异常名单的次数	

intangible_copyright.csv|icopy_num	企业软件著作权登记次数	
intangible_patent.csv|ipat_num	企业专利申请次数	
justice_credit.csv|is_justice_credit	列入失信黑名单的次数	
justice_credit_aic.csv|is_justice_creditaic	是否工商部失信企业	1：是；0：否
justice_declare.csv|appellant	企业作为原告的次数	
justice_declare.csv|defendant	企业作为被告的次数	
justice_enforced.csv|enforce_amount	企业执行标的	这里将同一家企业的所有执行标的进行了整合
justice_judge_new.csv|time	企业涉及到的法律纠纷次数	Time 是根据原表格数据提炼出的特征。简单来说，一家企业在这张表里出现了多少次，time 的值就是多少
trademark_infoa.csv|is_infoa	是否列为驰名商标	1：是；0：否
trademark_infob.csv|is_infob	列为著名商标的次数	

- 原始特征处理。有一些特征虽然对聚类结果帮助很大，但是特征本身存在一些别的问题，主要是格式问题，需要进行处理。如企业的注册资本 regcap 这一特征就存在这种问题。
- 特征合并

原始数据中还有一部分数据，需要组合起来才能达到我们的效果，这时就需要进行合并处理。如在处理 enterprise_insurance 这张表时，我们就统计了所有企业出现的次数（即企业给员工买的保险数目），并将结果赋予了 cbzt 这个特征，覆盖了它先前所代表的意义。

采用的特征中，若干个特征可以组合起来表征企业某一方面的特征。本项目中，我们为企业设计了五个方面的考量因素，分别是企业背景、企业经营能力、企业经营风险、企业司法风险、企业发展状况。表 1-5 详细说明了分别用来表征这五个方面的特征，其中有些特征出现在不止一个因素中。

表 1-5 特征分类表

考量因素	特征
企业背景	注册资本
企业经营能力	出质股权次数

	企业累计欠税额
	中标次数
	对外投资次数
企业经营风险	企业被列入经营异常的次数
	企业被行政处罚的次数（综合）
	出质股权次数
	企业累计欠税额
	企业被行政处罚的次数
	被列为守合同重信用企业的次数
	被列入异常名单的次数
	列入失信黑名单的次数
	是否工商部失信企业
企业司法风险	企业作为原告的次数
	企业作为被告的次数
	企业执行标的
	企业涉及到的法律纠纷次数
企业发展状况	企业给员工上的保险数目
	分支机构数
	网店个数
	企业软件著作权登记次数
	企业专利申请次数
	是否列为驰名商标
	列为著名商标的次数

- 空值填充

原始数据集中，很多需要用到的特征存在着部分缺失，这时就需要填充默认值，本项目中选用的默认填充值是 0。

本项目中主要的数据预处理工作就是上面四部分，另外需要注意的是，有一些特征在我们看来对企业的分类是有很大作用的，但本身数据过于稀少，如 trademark_infoa 表中的 is_infoa 特征。对于这类特征，我们采取的做法是仍然采用这些特征，而不是因为它们的稀疏就弃之不用。

(2) 数据归一化

经过数据预处理步骤得到的某些特征在取值范围上有比较大的差距，为了保证每个特征对目标函数的影响权重一致，需要进行归一化处理。归一化时，对每个数据做如下处理：

$$X^* = \frac{X - \min\ (D)}{\max(D) - std(D)}$$

其中，X 表示原始数据，D 表示 X 所在的特征所有数据组成的向量，$\min\ (\cdot)$、$\max\ (\cdot)$、$\text{std}\ (\cdot)$三个函数的作用分别是得到输入向量的最小值、最大值和标准差。

(3) 聚类

使用 Python 语言，借助 K-Means 算法，对处理后的数据进行聚类，我们分别得出类别数目为 2、3、4、5、6 的聚类结果。最后通过不同聚类数目下指标的变化情况，确定最好的聚类效果为 3 类。

(4) 类别标签标注

借助 Python 可视化技术，分别针对企业背景、企业司法风险、企业经营风险、企业经营能力等维度进行多类型图表分析，最终标注出每个类别的标签。

1.5.2. 指标评估

使用 Inertia、Calinski-Harabasz 指标和轮廓系数（Silhouette Coefficient）三个指标来综合评价每个聚类结果的好坏。效果如图 1-2 所示。

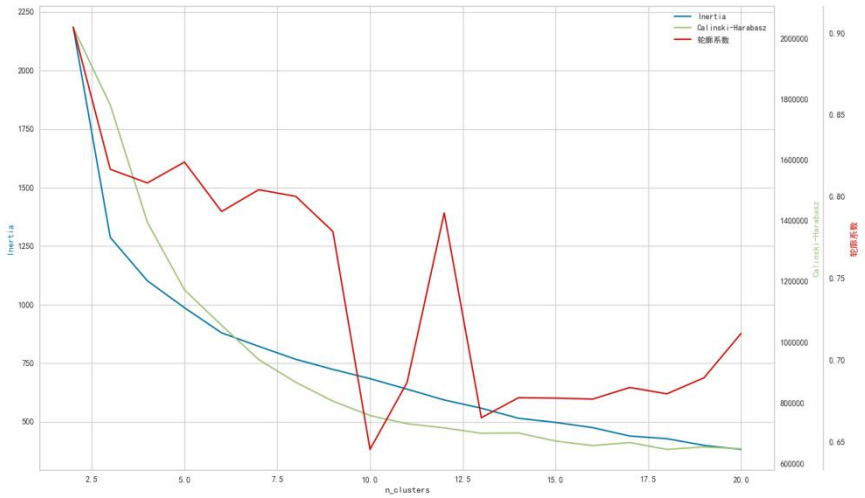

图 1-2 指标评估

1.5.3. 聚类指标可视分析

由于本项目的特征维度较高，不像二维或三维数据一样可以直观地展示聚类的分布情况，而若是将数据先降维再可视化的话容易损失很多细节，所以我们决定用雷达图、折线图、平行坐标图来可视化聚类结果。

（1）雷达图

雷达图非常善于展示多特征围度的数据，经过简单的处理后，雷达图的每条轴也可以表示各个不同聚类相应特征所占的比例。本项目中，即采用了这种方法，表示的是千分比。需要注意的是，由于展示一个聚类的某些特征时，采用的是计算总和的方法，所以企业数目更多的聚类有更大的概率在这些特征上取得更高的额占比。

图 1-3 是对聚类结果的总结性分析，每条轴所代表的企业信息即由特征计算加权和得来，具体情况是：

1）企业背景。只包括了注册资本一个特征。

2）企业经营能力$=-0.3*$出质股权次数$-0.2*$企业累计欠税额$+0.3*$中标次数$+0.2*$对外投资次数。

3）企业经营风险$=0.1*$企业被列入经营异常的次数$+0.05*$企业被行政处罚的次数(综合)$+0.05*$出质股权次数$+0.1*$企业累计欠税额$+0.05*$企业被行政处罚的次数$-0.15*$被列为守合同重信用企业的次数$+0.1*$被列入异常名单的次数$+0.2*$列入失信黑名单的次数$+0.2*$是否工商部失信企业。

4）企业司法风险=0.25 * *企业作为原告的次数* + 0.25 * *企业作为被告的次数* + 0.25 * *企业执行标的* + 0.25 * *企业涉及到的法律纠纷次数*。

5）企业发展状况=0.1 * *企业给员工上的保险数目* + 0.15 * *分支机构数* + 0.1 * *网店个数* + 0.15 * *企业软件著作权登记次数* + 0.15 * *企业专利申请次数* + 0.15 * *是否列为驰名商标* + 0.2 * *列为著名商标的次数*。

图 1-3 中，三个标签的小括号里的数字代表的是相应聚类中包含的企业数。从图中可以看出，在当前的评价体系下，聚类 0 的企业经营能力总体上最优，同时企业背景即注册资本虽最少但相差不大，企业经营风险也非常低，但需要注意的是，聚类 0 的企业司法风险是最高的，企业发展状况也不是太好；聚类 1 在企业背景、企业发展状况两方面都占据优势，特别是企业发展状况方面，同时企业经营能力稍逊于聚类 0，有一定的企业经营风险和企业司法风险，需要注意；聚类 2 的企业比较特殊，企业背景方面与其他聚类相差不大，但企业经营风险很高，而在企业发展状况、企业司法风险和企业经营能力方面都只占据了一个很小的比例。

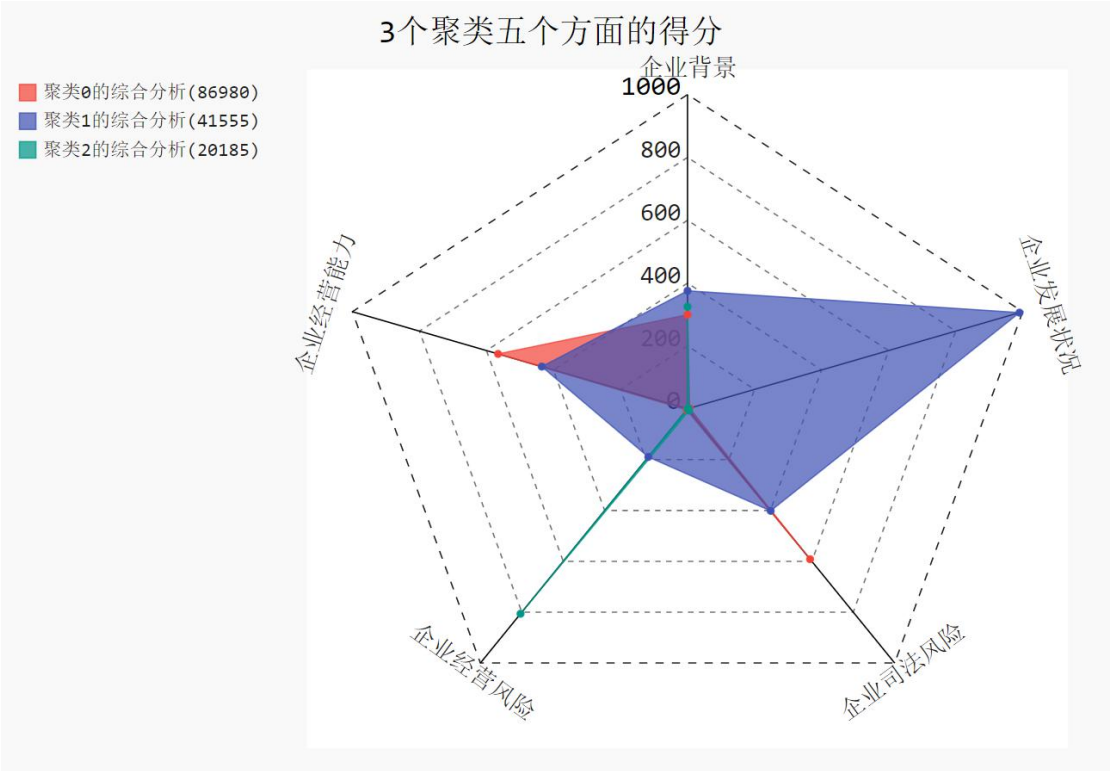

图 1-3 3 个聚类 5 个方面的得分

- 雷达图

图 1-4~1-7 分别用雷达图展示了企业的发展状况、司法风险、经营风险、经营能力等，这些方面的情况都用若干个特征表示。

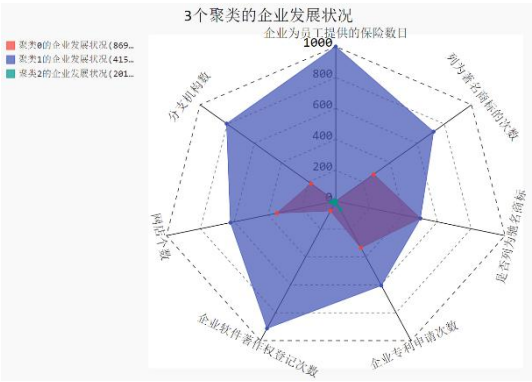

图 1-4 企业发展状况

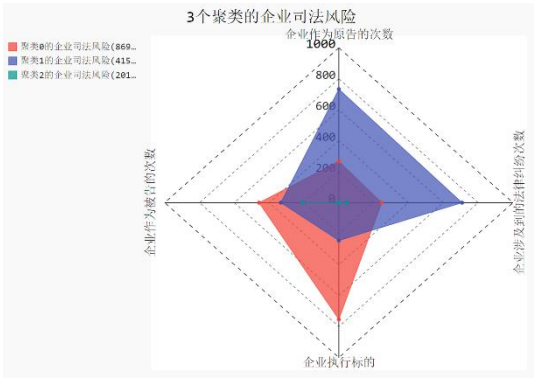

图 1-5 企业司法风险

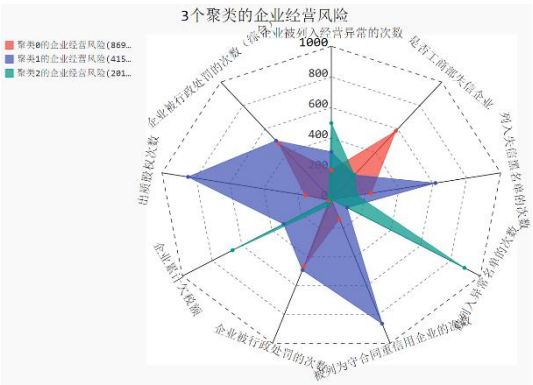

图 1-6 企业经营风险

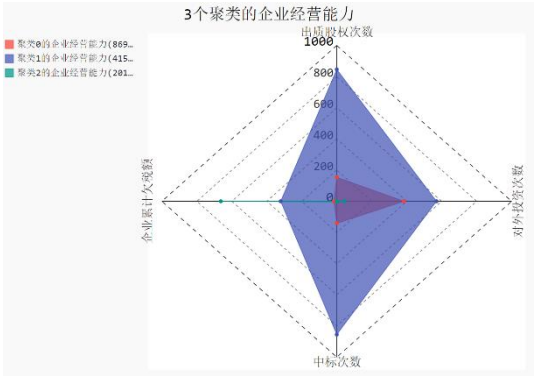

图 1-7 企业经营能力

图 1-8~-1-11 展示了企业背景里的一些信息，需要注意的是，除了注册资本外，其他的信息并没有参与之前的聚类，而只是用来进行展示。图 1-11 中，我们对注册资本做了不同层次的划分，但这种划分是不均匀的，因为在 0-500 这个范围内聚集了大量的企业，需要进行更细致的划分

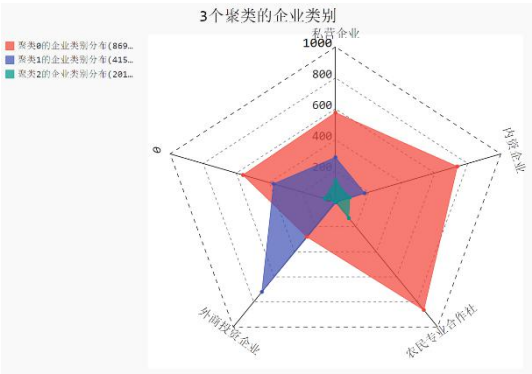

图 1-8 企业背景-企业类别

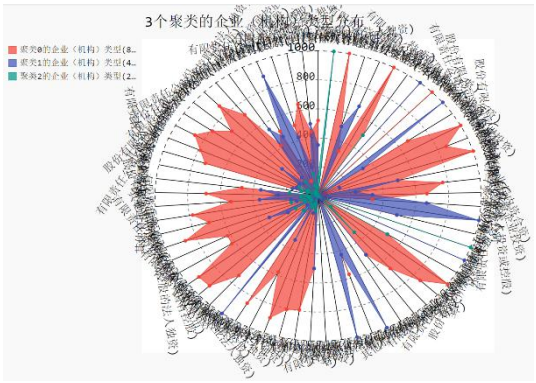

图 1-9 企业背景-企业（机构）类型

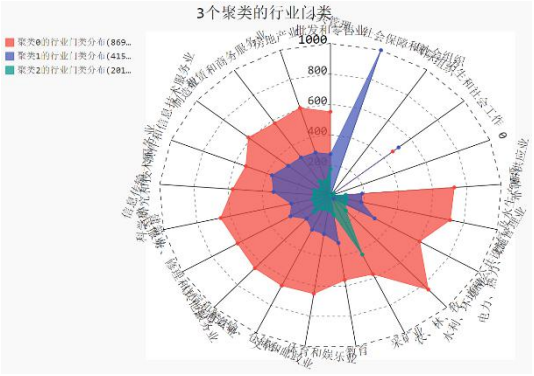

图 1-10 企业背景-行业门类

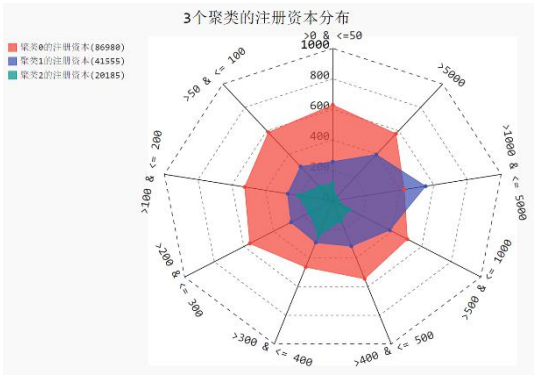

图 1-11 企业背景-注册资本

- **折线图**

折线图也是一种可以用来可视化多维度特征数据的较好办法。在绘制折线图时，我们并没有全部采用千分比的形式来绘制，大部分是原始数据，或是放大以更好可视化的原始数据，只有小部分数据由于数据值太大会影响其他数据的展示而采用了千分比的形式。

图 1-12~1-15 展示了三个聚类在企业发展状况、企业司法风险、企业经营风险、企业经营能力四个方面的表现。图中，X 轴有些特征后面带有小括号，小括号里面若是“‰”，说明这个特征的 Y 轴表示的是千分比；若是“*数字”，则说明在可视化时将原始数据扩大了相应倍数，这是因为不同特征的量纲有时会差距过大。

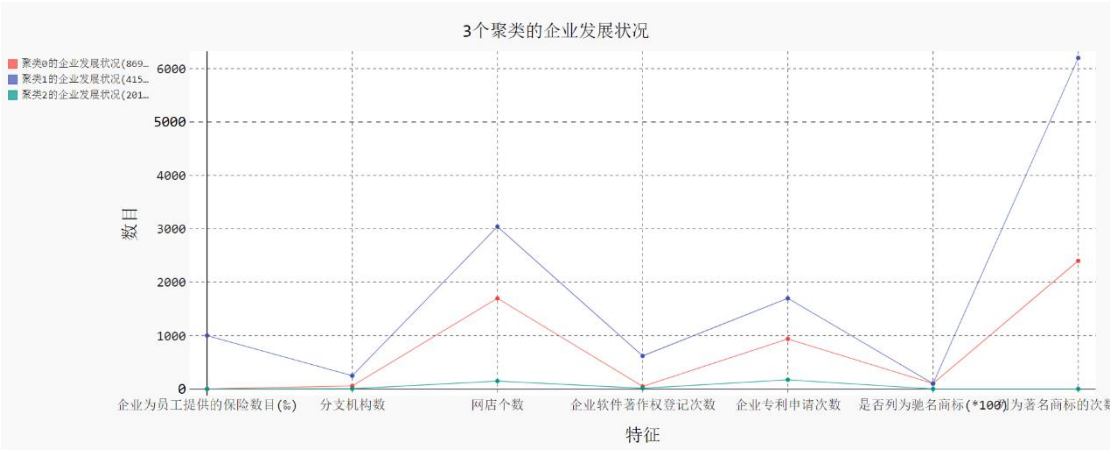

图 1-12 企业发展状况

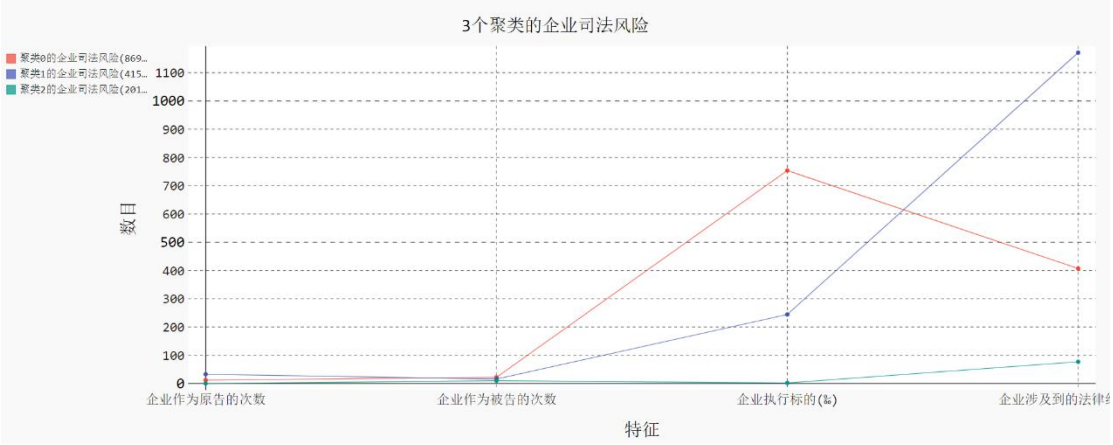

图 1-13 企业司法风险

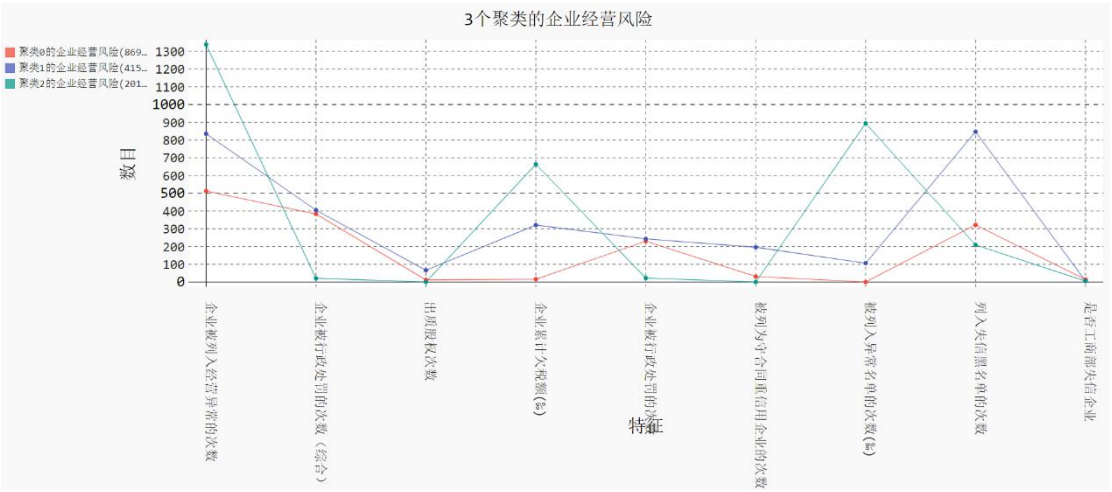

图 1-14 企业经营风险

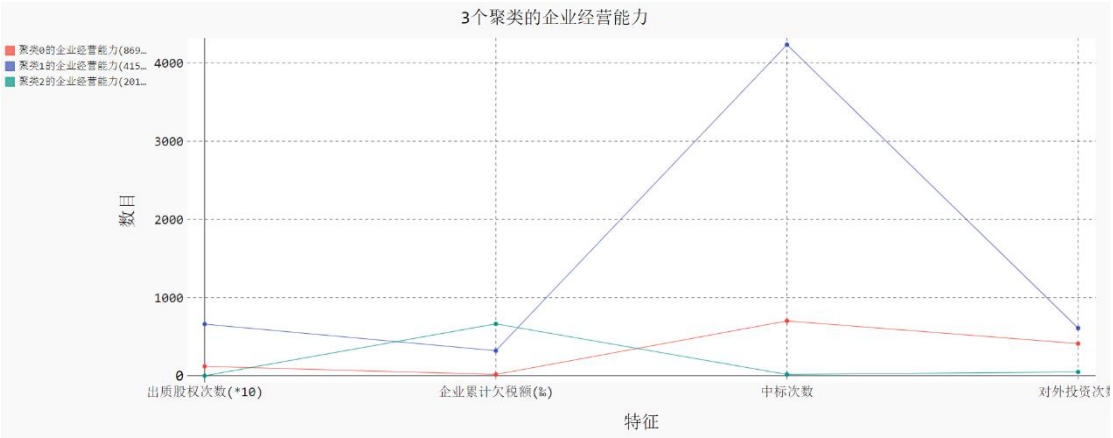

图 1-15 企业经营能力

图 1-16~1-19 展示了企业背景方面的一些信息，图中，X 轴的某些特征是“0”，这说明这部分企业缺失了该特征，于是填充了默认值。从图中可以看出，大部分企业是私营企业，并且企业（机构）类型多是有限责任公司（自然人投资或控股）或有限责任公司（自然人独资）；企业的行业门类最多的则是批发和零售业，租赁和商务服务业次之；注册资本方面，绝大部分在 500 万元以下。体现出来的这些特点很好地符合了中小企业这个主体。

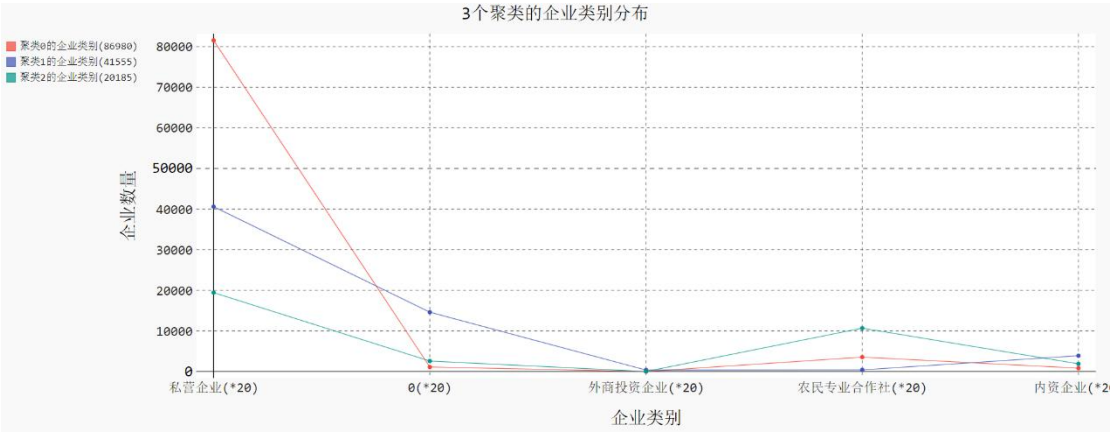

图 1-16 企业类别

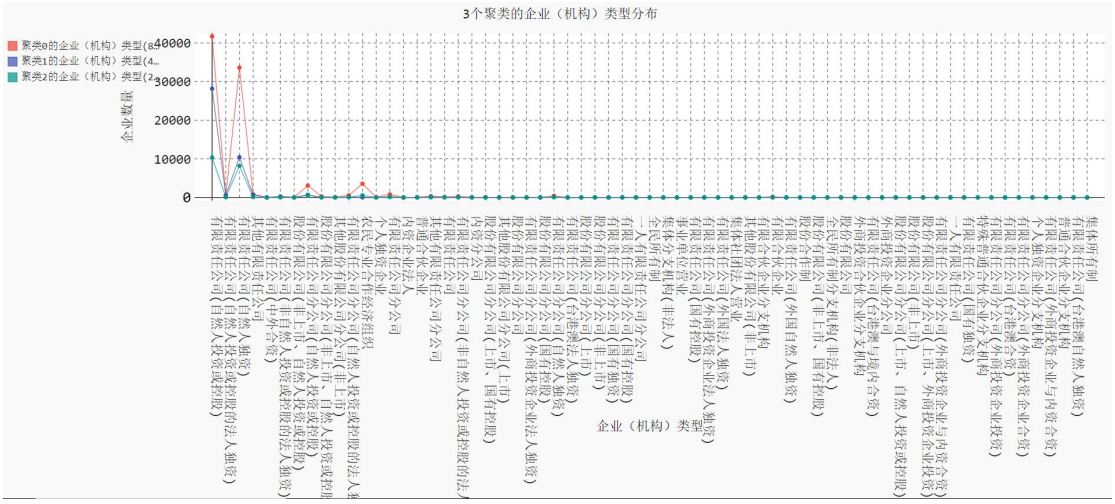

图 1-17 企业（机构）类型

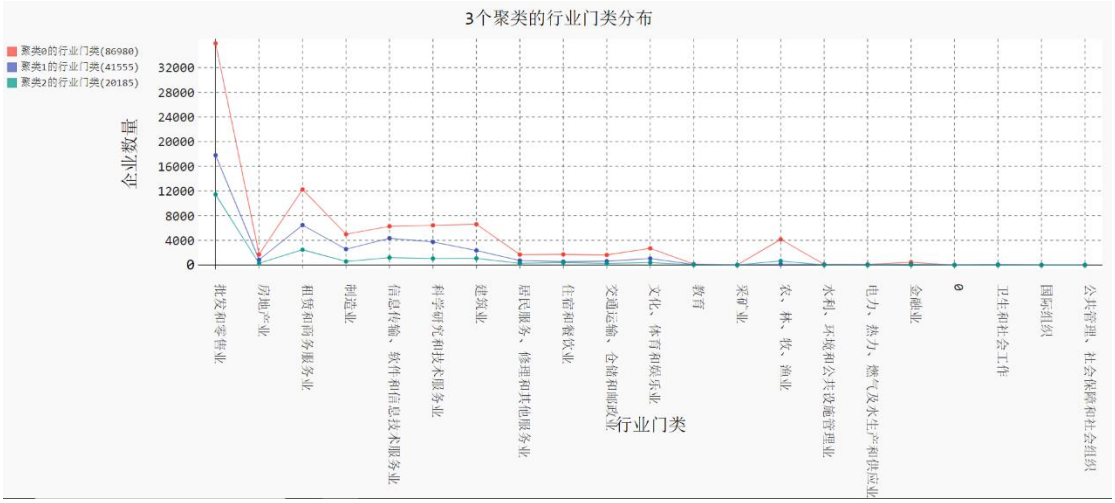

图 1-18 行业门类

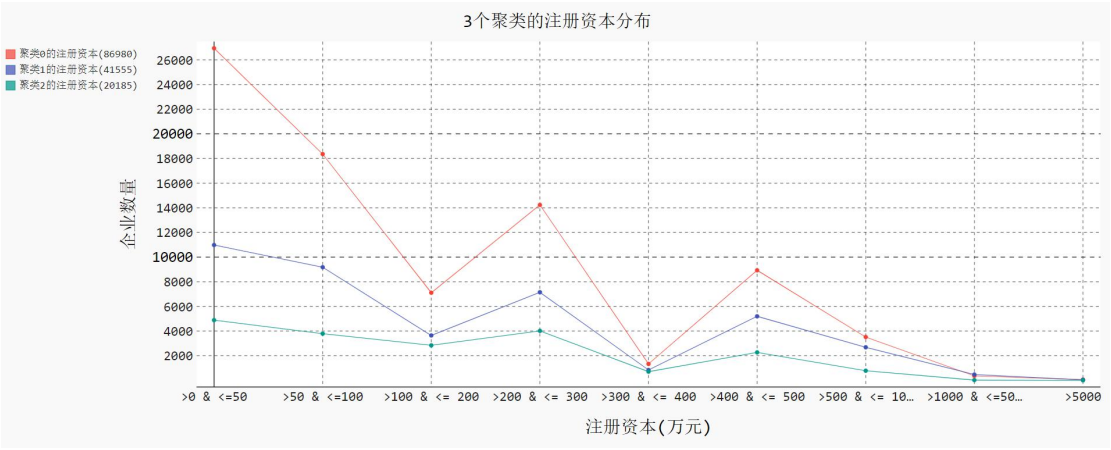

图 1-19 注册资本

- **平行坐标图**

平行坐标图在一张图上将一个样本的若干特征用折线连起来，并根据其所属的聚类标示颜色。相较而言，平行坐标图更容易碰到不同特征之间量纲不同的问题，因为有时候在一张平行坐标图上展示的特征数量是很大的，因此我们在绘制平行坐标图时，对数据进行了标准化处理。

图 1-20~1-24 用平行坐标图展示了企业的发展状况、司法风险、经营风险、经营能力、背景等五个方面，图 1-25 则将所有的特征都进行了展示。从这些图中可以看出，就本项目而言，平行坐标图所能体现的深层信息其实是比较少的，因此，它在本项目的聚类结果分析中更多承担的是辅助的作用。

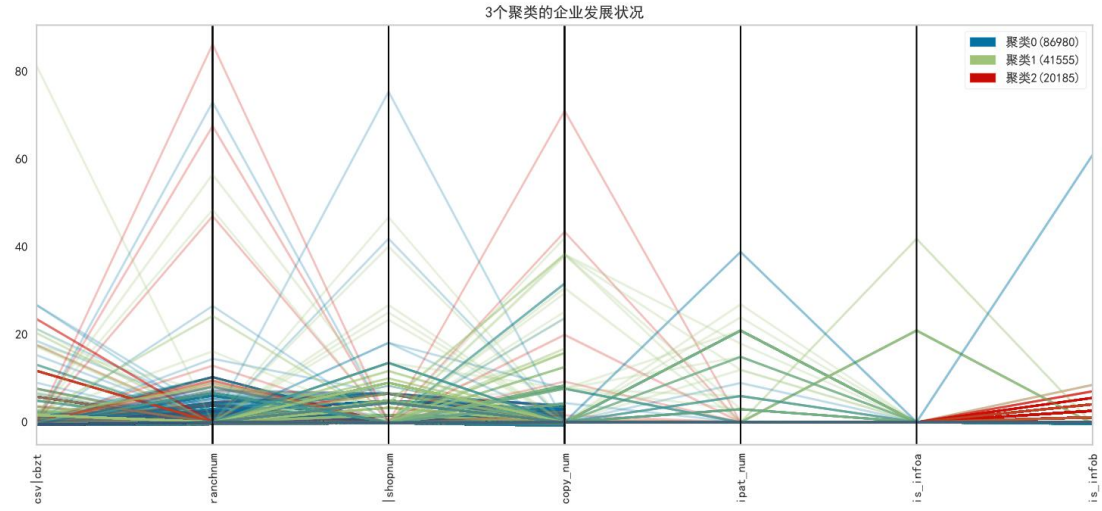

图 1-20 企业发展状况

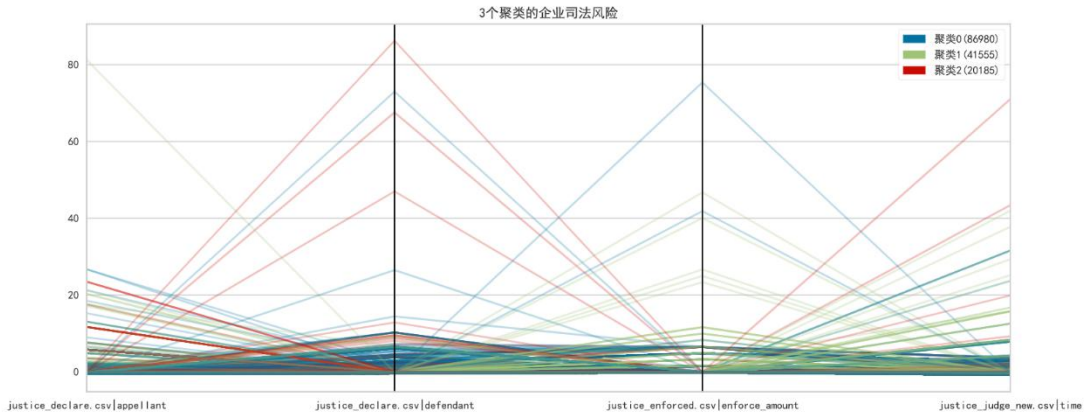

图 1-21 企业司法风险

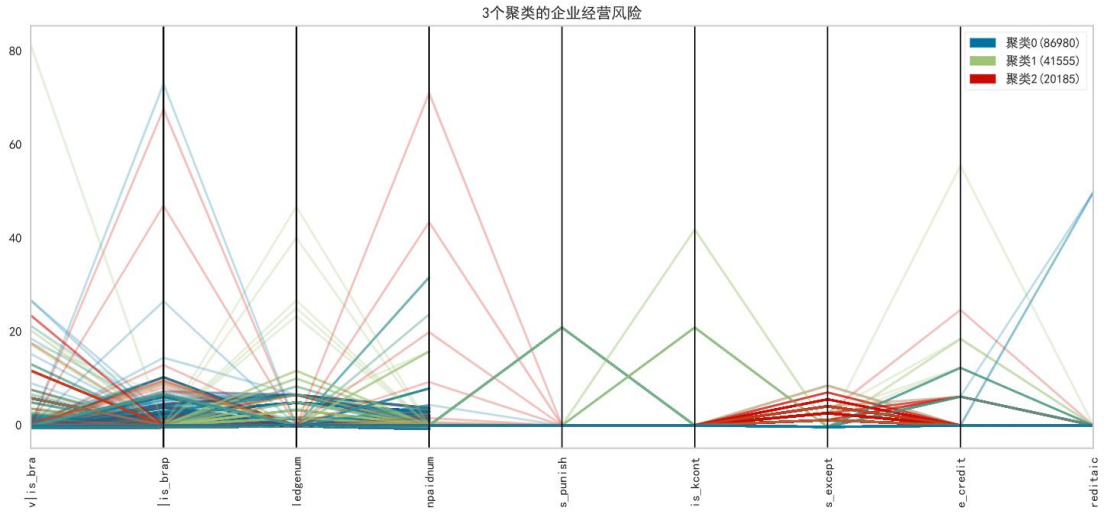

图 1-22 企业经营风险

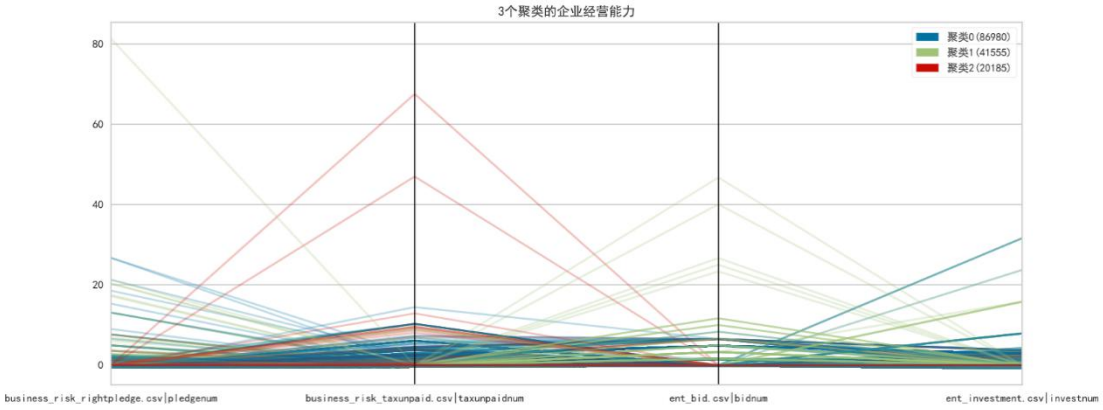

图 1-23 企业经营能力

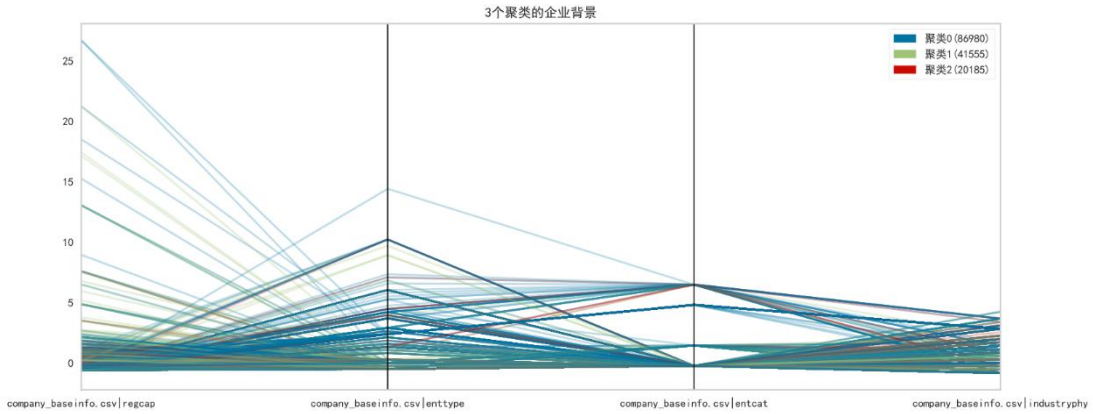

图 1-24 企业背景

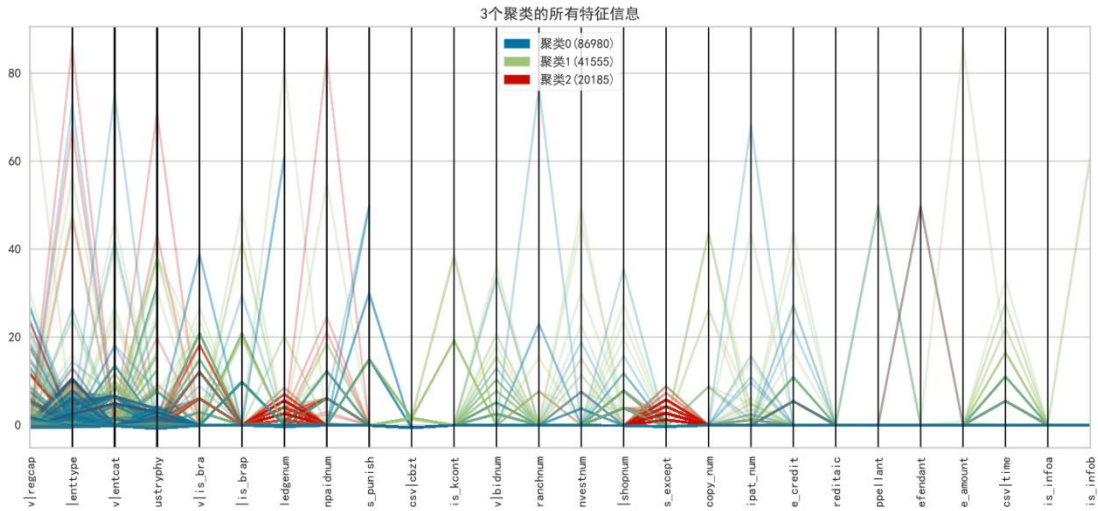

图 1-25 企业所有特征信息

1.5.4. Web 端可视化

（1）总视图

采用多视图协调关联，不同的试图模块展示企业数据的一部分属性。

图 1-26 总视图

（2）TSNE 散点图

我们对数据进行降维，考虑到本系统中的数据特征相关度并不高，直接采用线性方法不能较好地保留数据信息，因此考虑使用流形学习方法。TSNE 改变了 MDS 和 ISOMAP 中基于距离不变的思想，将高维映射到低维的同时，尽量保证相互之间的分布概率不变，SNE 将高维和低维中的样本分布都看作高斯分布，而 TSNE 将低维中的坐标当做 T 分布，这样做的好处是为了让距离大的簇之间距离拉大，从而解决了拥挤问题。并用 K-Means 进行聚类处理后，将得到的数据在 Java 端进行读取，并且在前端用 D3 进行展示。首先降维结果作为整体数据的 Overview，我们使用了不同的颜色表示不同公司类别（即聚类结果），可以直观看到数据在二维平面上的大致分布和相对位置。同时通过判断数据点的位置分布和颜色属性是否一致也可以验证聚类的效果为了更加突出每一类的效果，我们在散点图的右上角的小图例添加了鼠标点击功能，点击某一类别，散点图中该类别高亮，其他类别减小透明度。我们也在散点图中给每一个点添加了鼠标移入移出事件，鼠标移入某个点，与该点同一类别的点全部高亮，如图 1-27。

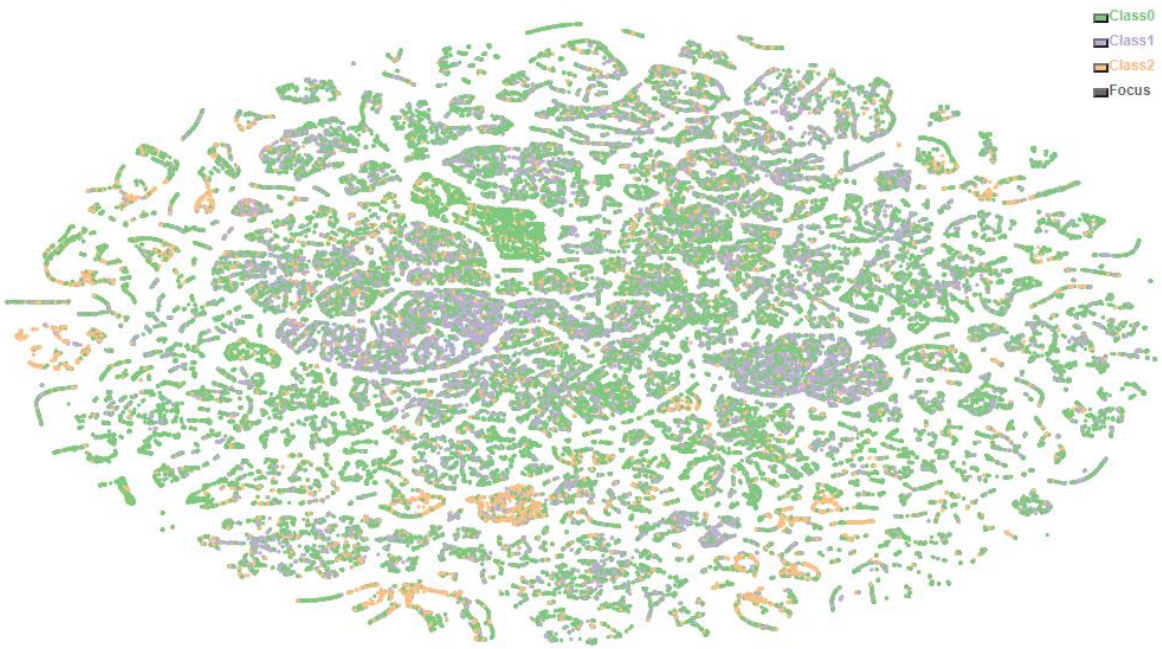

图 1-27 TSNE 散点图

在复杂的图形上同时观察微观和宏观特征可能很困难。如果放大以获得更多细节，则该图太大而无法完整查看。如果缩小以查看整体结构，则会丢失一些细节。为了更好的提升用户体验，我们还在散点图中添加了鱼眼效果，如图 1-28。

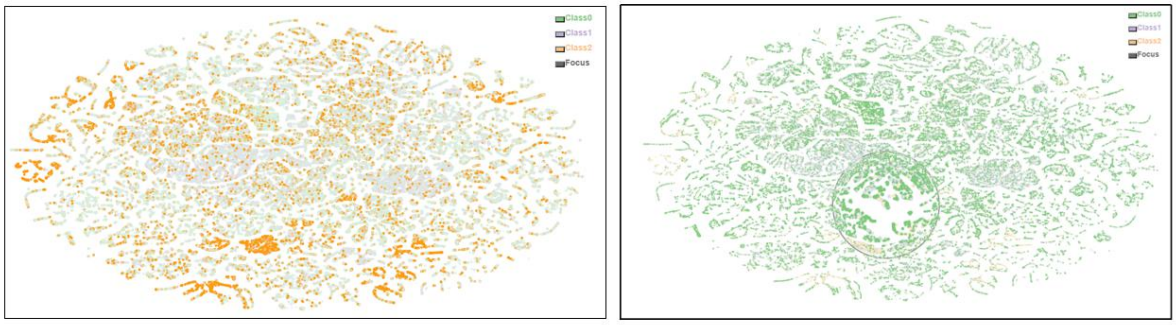

图 1-28 鱼眼、类别聚焦图

（3）类别标签词云图

力导向布局是建立在粒子物理理论的基础上，将节点模拟成为原子，通过原子间的引力和斥力来得到节点的速度与加速度，计算其移动方位与距离，最终达到一个稳定平衡的状态，从而完成布局。在 D3 的实现中，为了达到性能与效果的平衡，节点与节点间模拟同种电荷相互排斥，并将节点存入四叉树中，利用 Barnes－Hut 近似来减少节点间电荷斥力的计算量。同时连线间的节点模拟弹簧牵引力，节点的速度综合斥力引力得出，并发生阻尼衰减，最终达到整图平衡。

在标签词云图上加上力导向布局不仅保留了词云图将高频词突出显示的优点，也极大提高了用户的自由度。当词云图的布局不符合期望时，用户可自由的对每个词云节点

进行拖动，将想要展示的重点信息放在图形的主要位置，对原本的词云图进行想要的布局，让词云图更具灵活性和直观性。此外，对于每个词云，我们都绑定了鼠标事件，当鼠标移入词云时，词云颜色加深并且放大、移动，以突出显示。如图 1-28 自定义词云图，重点标签如“外商投资”、“守合同信用”等处于词云图中间位置更加突出，鼠标悬停在某标签处，此标签颜色加深并且放大，如图 1-29。

图 1-28 自定义词云图

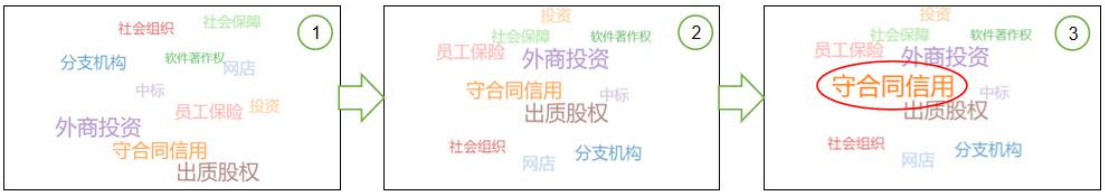

图 1-29 交互词云图

（4）评分雷达图

雷达图是以从同一点开始的轴上表示的三个或更多个定量变量的二维图表的形式显示多变量数据的图形方法。轴的相对位置和角度通常是无信息的。 雷达图也称为网络图，蜘蛛图，星图，蜘蛛网图，不规则多边形，极坐标图或 Kiviat 图。它相当于平行坐标图，轴径向排列。雷达图主要应用于企业经营状况——收益性、生产性、流动性、安全性和成长性的评价。上述指标的分布组合在一起非常象雷达的形状，因此而得名。

我们采用 ECharts 雷达图对聚类结果进行总结性分析，每条轴代表的企业信息由对应特征计算加权和得来，分别为企业背景、企业经营能力、企业经营风险、企业司法风

险、企业发展状况。如图 1-30 所示，三个标签的小括号里的数字代表的是相应聚类中包含的企业数。绿色代表聚类 0，紫色代表聚类 1，黄色代表聚类 2。从雷达图中我们得知，聚类 0 的企业背景虽低但是分布较为均匀，同时企业经营风险也很低，但同时我们也发现，聚类 0 的企业司法风险很高的，企业发展状况也不是太好；聚类 1 在企业背景、企业发展状况、企业经营能力等方面都表现不错，但也有一定的企业经营风险和企业司法风险；聚类 2 的企业背景方面与其他类相差不大，但企业经营风险很高，而在企业司法风险和企业经营能力方面都只占据了很小的比例。

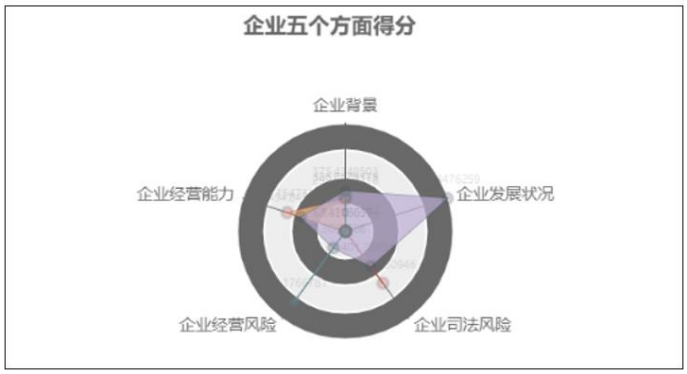

图 1-30 评分雷达图

（5）检索与企业信息、类别展示

如图 1-31 支持企业名称及相关特征检索，并在展示模块显示企业的基本信息和类别信息，如图 1-32。

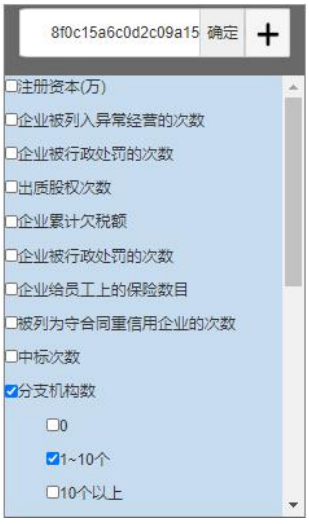

图 1-31 企业检索

sort	企业名称	注册资本	从业人数	成立日期	企业（机构）类型	企业类别	行业门类	经营（驻在）期限自	企业被列入经营异常的次数	企业被行政
	49c10409ee98ceac53d5eb1338b22e5b	51	8	2008年03月24日	1	1	1	2008年03月24日	0	
	fab957c0a496f2b486c0c423ffd7488b	288	8	2008年03月24日	3	1	4	2008年03月24日	1	
	af34ceff0caf0aed82d082393313fd2e	51	8	2008年03月24日	1	1	5	2008年03月24日	0	
	77a80667ce5ff28d2cb4c5525a136a18	51	8	2008年03月24日	1	1	6	2008年03月24日	0	
	2a416404a54360441f6155a1ac2b4295	51	8	2008年03月24日	1	1	1	2008年03月24日	0	
	5f80de57aa4fdc76c4001824ab8fa8f6	500	8	2008年04月16日	1	1	1	2008年04月16日	0	

图 1-32 企业信息及类别展示

1.6.数据库设计

本项目针对处理后的数据进行数据库设计，如图 1-33 所示：

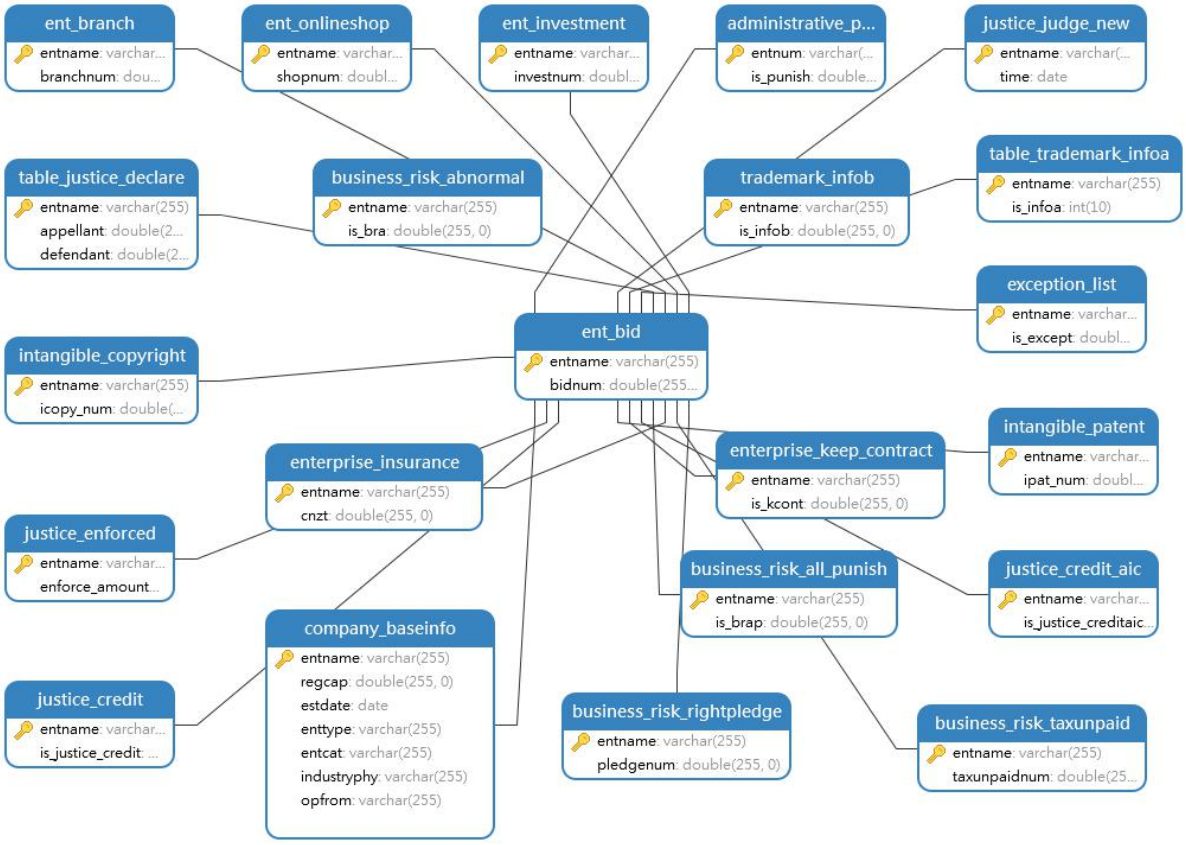

图 1-33 数据库设计

详细数据库表设计如下所示：

表 1-6 企业基本信息表

company_baseinfo（基本信息）		
字段	规格	描述
regcap	Double(255)	注册资本

empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
esdate	Date(14)	成立日期
enttype	Varchar(255)	企业（机构）类型
entcat	Varchar(255)	企业类别
industryphy	Varchar(255)	行业门类
opfrom	Varchar(255)	经营（驻在）期限自

表 1-7 企业经营风险-经营异常表

business_risk_abnormal（经营异常信息）		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
is_bra	Double(255)	企业被列入经营异常的次数

表 1-8 企业经营风险-行政处罚表

business_risk_all_punish(行政处罚信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
is_brap	Double(255)	企业被列入经营异常的次数

表 1-9 企业经营风险-股权出质表

business_risk_rightpledge(股权出质信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
pledgenum	Double(255)	出质股权次数

表 1-10 企业经营风险-欠税公告

business_risk_taxunpaid(欠税公告信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
taxunpaidnum	Double(255)	企业累计欠税额

表 1-11 行政处罚表

administrative_punishment(行政处罚信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），

		唯一标识一个企业
is_punish	Double(255)	企业被行政处罚的次数

表 1-12 单位参保信息查询（养老单位参保信息）

enterprise_insurance(参保信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
cbzt	Double(255)	企业给员工上的保险数目

表 1-13 守合同重信用企业

enterprise_keep_contract(守合同重信用信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
is_kcont	Double(255)	被列为守合同重信用企业的次数

表 1-14 企业中标表

ent_bid(中标信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
bidnum	Double(255)	中标次数

表 1-15 企业分支机构信息

ent_branch(分支机构信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
|branchnum	Double(255)	分支机构数

表 1-16 企年报对外投资表

ent_investment(对外投资信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
investnum	Double(255)	对外投资次数

表 1-17 年报网店信息表

ent_onlineshop(网店信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
shopnum	Double(255)	网店个数

表 1-18 异常名单表

exception_list(异常名单信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
is_except	Double(255)	被列入异常名单的次数

表 1-19 知识产权_软件著作权数据表

intangible_copyright(软件著作信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
icopy_num	Double(255)	企业软件著作权登记次数

表 1-20 知识产权_专利数据

intangible_patent(专利信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
ipat_num	Double(255)	企业专利申请次数

表 1-21 司法风险-失信黑名单数据表

justice_credit(黑名单数据信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
is_justice_credit	Double(255)	列入失信黑名单的次数

表 1-22 失信企业（工商部）表

justice_credit_aic(失信企业信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
is_justice_creditaic	Integer(10)	是否工商部失信企业

表 1-23 司法风险—开庭公告数据表

justice_declare(开庭公告数据信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
appellant	Double(255)	企业作为原告的次数
defendant	Double(255)	企业作为被告的次数

表 1-24 司法风险—被执行人数据表

justice_enforced(被执行人数据信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
enforce_amount	Double(255)	企业执行标的

表 1-25 司法风险-裁判文书数据表

justice_judge_new(裁判文书信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
time	Double(255)	企业涉及到的法律纠纷次数

表 1-26 驰名商标信息表

trademark_infoa(驰名商标信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
is_infoa	Integer(10)	是否列为驰名商标

表 1-27 著名商标信息信息表

trademark_info(著名商标信息)		
字段	规格	描述
empnum	Varchar(255)	企业名称（加密处理），唯一标识一个企业
is_infob	Double(255)	列为著名商标的次数