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 无监督分类方法与参数调优.....3
1.1 K-Means 聚类.....3
1.2 参数调优.....3
2 训练结果与簇划分.....5
2.1 模型训练结果.....5
2.2.簇划分.....5
2.3 聚类结果分析与标签标注.....7
3 Web 前端可视化.....18
3.1 单个企业检索.....18
3.2 批量检索.....18
3.3 TSNE 降维可视化.....19
3.4 标签词云可视化.....20
3.5 雷达图表分析.....20

1 无监督分类方法与参数调优

1.1 K-Means 聚类

K-Means 是无监督聚类领域比较基础，同时因为不错的聚类效果，应用范围也非常广泛的一种算法。它的思想较为简单，根据样本间的距离大小，将数据分为预先设定的 K 个聚类。

K-Means 主要可分为以下几个步骤。

1）K 的确定。一般来说，K 的确定是根据先验经验来选择的，如果没有丰富的先验知识，可以选取若干个 K 值跑以下算法，根据后续的评价指标来确定最终的 K 值。

2）选择质心。确定 K 的值后，需要选定 K 个质心来开始算法的迭代，这 K 个质心要尽可能的远一点。质心的选择对后续的聚类结果和聚类时间都有一定程度的影响。

3）聚类。有了前两步的铺垫后，算法就可以进行迭代以进行聚类了。

k-means 聚类中除了质心的选择外，还有一个重要的地方就是距离度量的方式，本项目尝试了两种方法，余弦距离和欧几里得距离。

余弦距离是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的度量。设有两个 n 维向量和，则这两个向量的夹角的余弦值可由下列公示得到：

欧几里得距离计算的则是二维坐标系里两点之间的距离在 n 维特征空间中的扩展，具体地，对 n 维向量和来说，它们之间的欧几里得距离为：

本项目中，使用的是欧几里得距离。

1.2 参数调优

对聚类任务而言，参数调优的目的就是使聚类出的结果尽量的“物以类聚”，即簇内相似度尽可能高，同时簇间相似度又要尽可能得高。而要评价这种相似度，就需要用到某种指标。一般而言，用来评价相似度的指标可以大体分为两类，一类将聚类结果与某个“参考模型”比较，称之为外部指标；另一类则直接考察结果，称为内部指标。

外部指标所用的参考模型，某种意义上代表的就是“正确”的聚类模型。对某个样本 i 而言，令C_i表示聚类模型中 i 的标记，C_i^*表示参考模型中 i 的标记，则若将样本两两配对考虑，对样本x_i, x_j的聚类会有四种情况：1）$C_i = C_j, C_i^* = C_j^*$；2）$C_i = C_j$，$C_i^* \neq C_j^*$；3）$C_i \neq C_j$，$C_i^* = C_j^*$；4）$C_i \neq C_j$，$C_i^* \neq C_j^*$。其中第一种情况是x_i和x_j在聚类模型中属于同一类，同时在参考模型中也是同一类，其他三种则是其他的可能。假设数据集中属于上述四种情况的样本数分别为 a，b，c，d，则可借此给出三种外部指标的定义。

1）Jaccard 系数：

$$JC = \frac{a}{a+b+c}$$

2）FM 指数：

$$FMI = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

3）Rand 指数：

$$RI = \frac{2(a+d)}{m(m-1)}$$

上述三种外部指标的值都在[0,1]区间，并且越接近 1 说明效果越好。

对于本项目而言，由于没有“参考模型”，因此用来评估模型好坏时，只能采用内部指标。具体来说，本项目采用了以下三个内部指标来选择最佳的聚类数目。

1）Inertia。这个指标计算的是聚类中各个样本到聚类中心的距离之和，inertia 越小，说明聚类越紧密。

2）Calinski-Harabasz。CH 指标的计算相对来说更加复杂：

$$CH(k) = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{m-k}{k-1}$$

其中，m 为训练集样本数，k 为类别数，B_k为聚类之间的协方差矩阵，W_k 为聚类内部的协方差矩阵，tr 为矩阵的迹。CH 分数越高则聚类效果越好，即类内协方差越小越好，类间协方差越大越好。

3）Silhouette Coefficient，即轮廓系数。轮廓系数的计算方式也比较直接：

$$S = \frac{b(i) - a(i)}{\max \quad \{a(i), b(i)\}}$$

其中，i 指的是样本点，$a(i)$为 i 与同聚类内所有其他样本的平均距离。对除了 i 所在的聚类外的所有聚类，计算 i 与它们所包含的所有样本的平均距离，其中最小的平均距离就记为 $b(i)$。可以看出，轮廓系数的取值范围是[-1, 1]，并且越靠近 1，说明聚类效果越好。

下图 1-1 展示了聚类个数分别取[2, 20]范围内的不同值时，三个指标的得分情况。在选取最佳聚类数目时，通常会采用“肘部法则”，即选取曲线拐点处的聚类数作为最佳聚类数目。在图中，综合三个指标来看，可以看出 $n\_clusters = 3$ 时 inertia 和轮廓系数曲线都出现了拐点，同时 Calinski-Harabasz 系数的值也处在一个非常能够接受的位置，因此我们选定最佳聚类数目为 3。

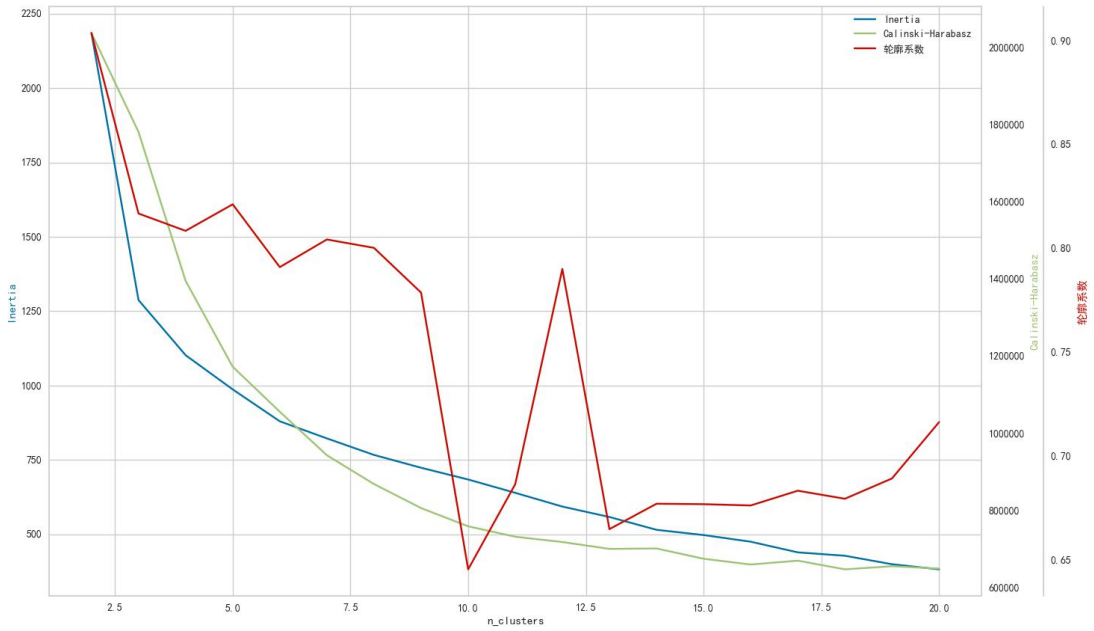

图 1-1 不同聚类数目下三种指标的变化情况

2 训练结果与簇划分

2.1 模型训练结果

本项目中在存储聚类模型时采用的是 sklearn 中的 joblib 模块。见补充材料 K-Means.model。

2.2.簇划分

如图 1-1 展示了聚类个数分别取[2, 20]范围内的不同值时，三个指标的得分情况。

我们又尝试使用 DBSCAN 实现聚类，如图所示，效果并不明显。

后来我们针对聚类为 3 类、4 类、5 类的结果分别进行 TSNE 降维可视化，发现当簇的数目增加到 4 类、5 类时，聚类结果并没有更明显的划分。

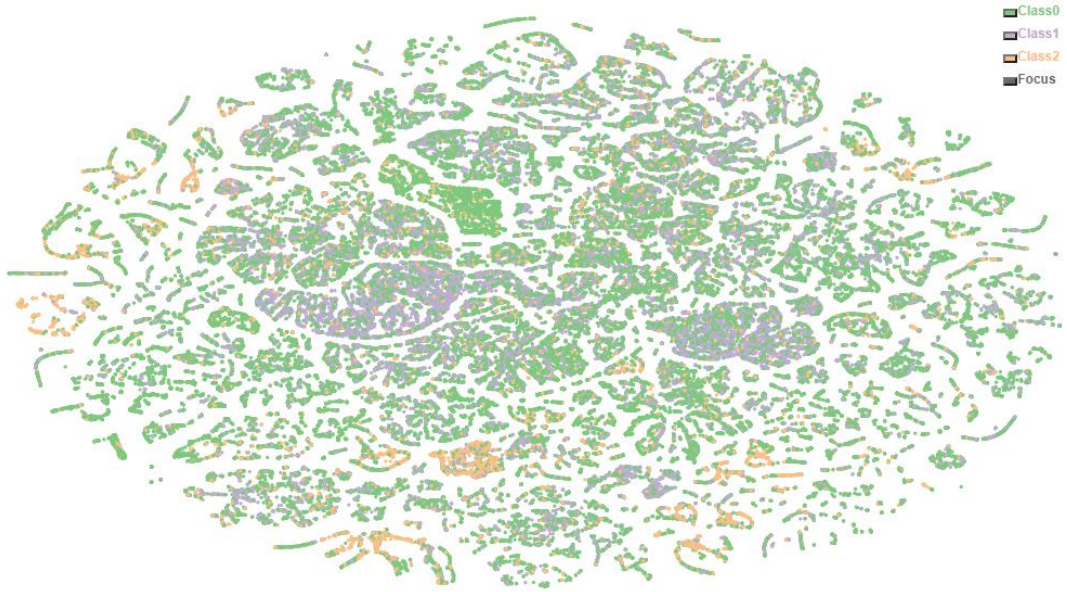

图 2-1　3 类聚类 TSNE 降维结果

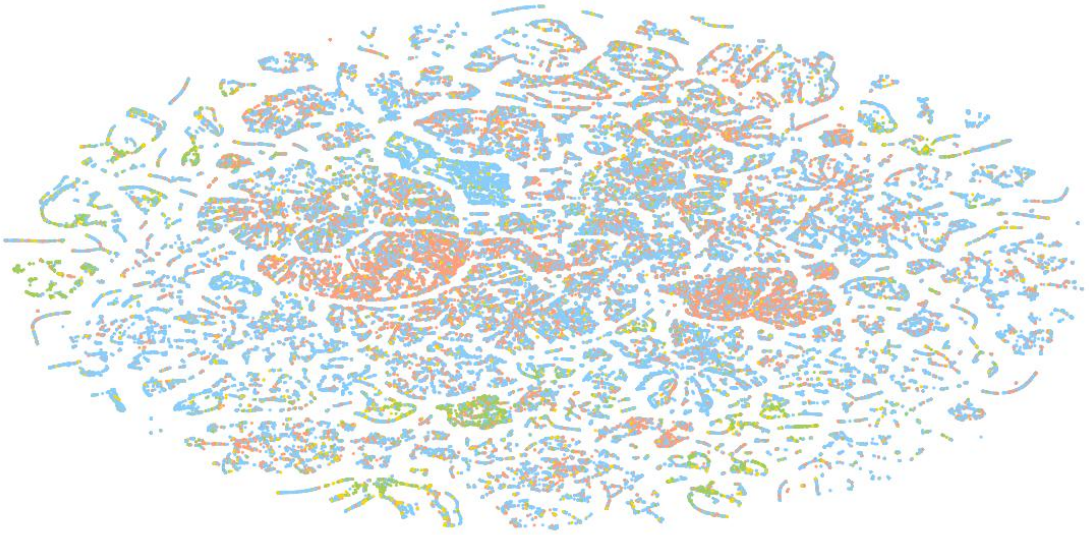

图 2-2　4 类聚类 TSNE 降维结果

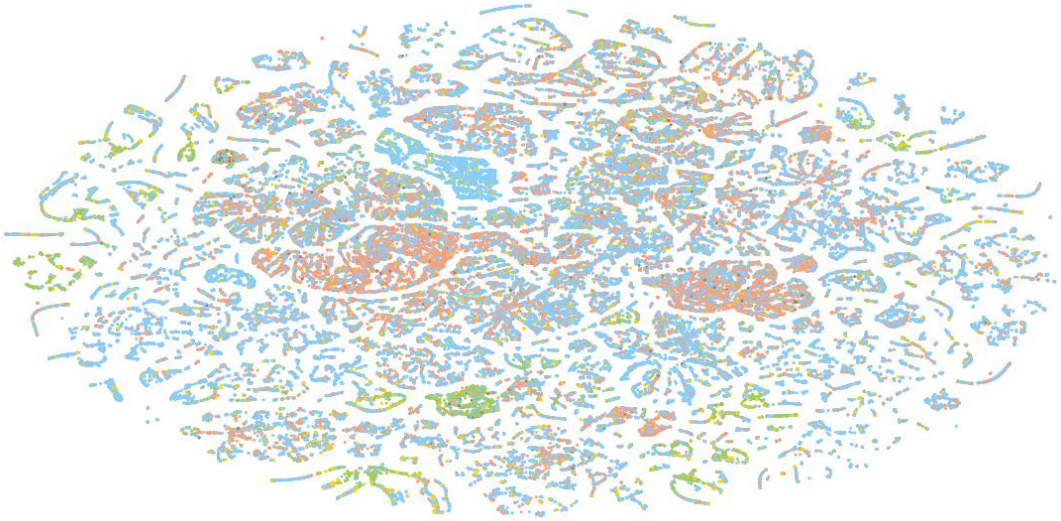

图 2-3　5 类聚类 TSNE 降维结果

综合这两方面我们可以得出最好的聚类数目是 3。

2.3 聚类结果分析与标签标注

由于本项目的特征围度较高，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。

- 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 2-4 是对聚类结果的总结性分析，每条轴所代表的企业信息的对应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 2-4 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。

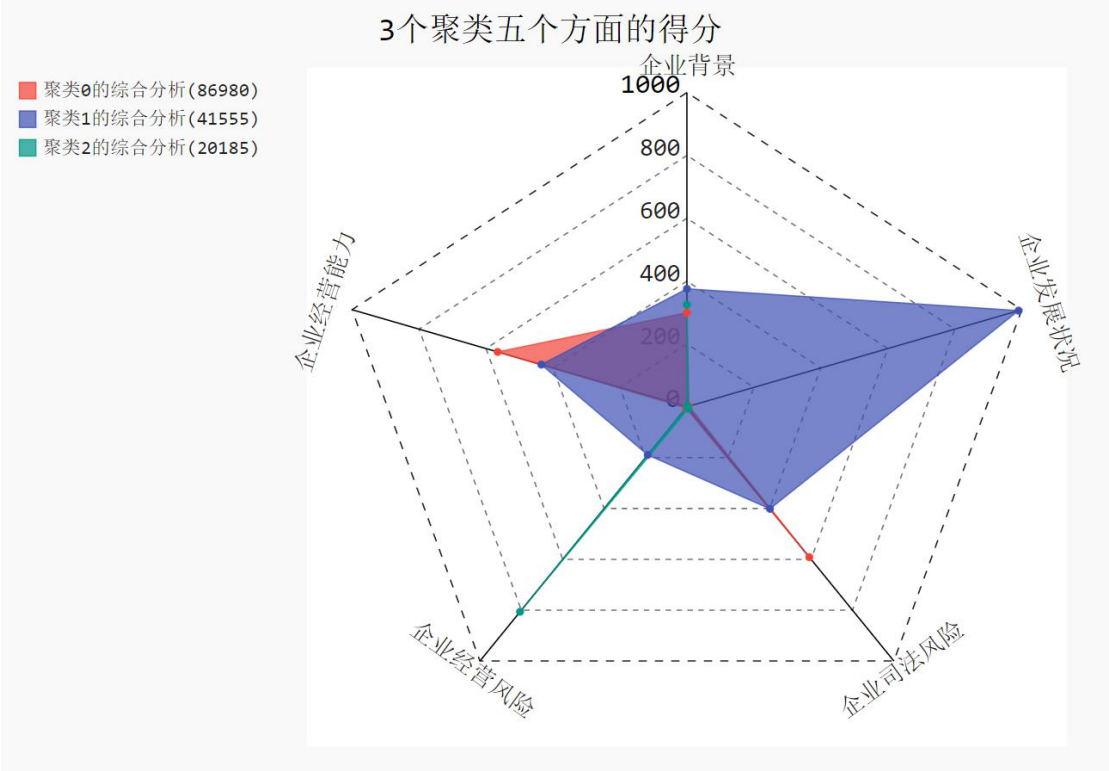

图 2-4　3 个聚类 5 个方面的得分

图 2-5~2-8 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。

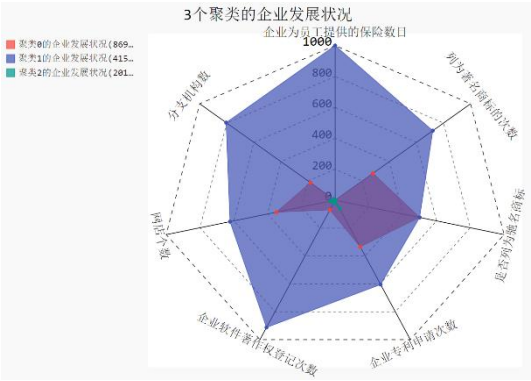

图 2-5　企业发展状况

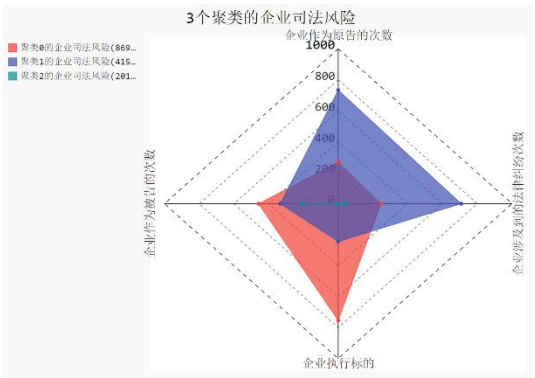

图 2-6 企业司法风险

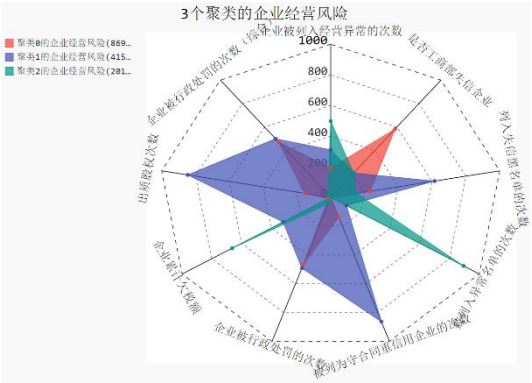

图 2-7　企业经营风险

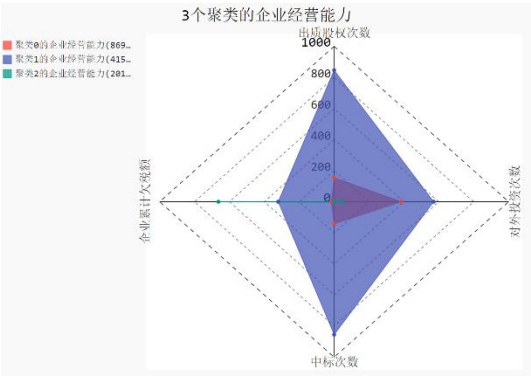

图 2-8　企业经营能力

图 2-9~2-12 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 2-12 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

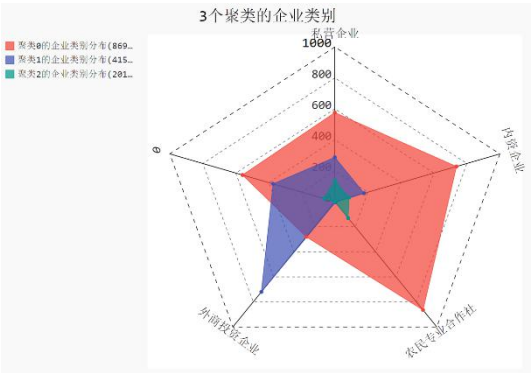

图 2-9　企业背景-企业类别

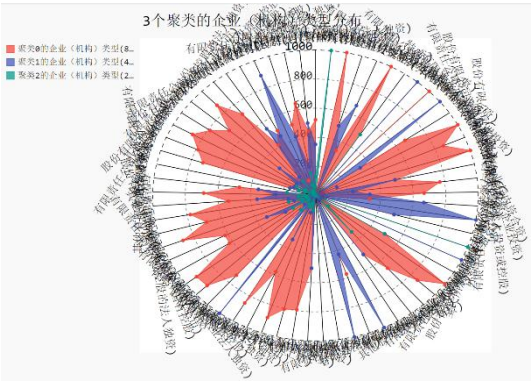

图 2-10 企业背景-企业（机构）类型

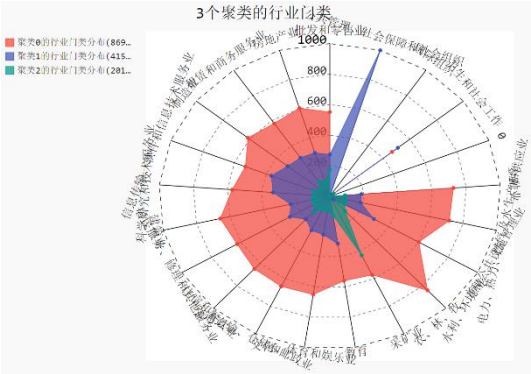

图 2-11　企业背景-行业门类

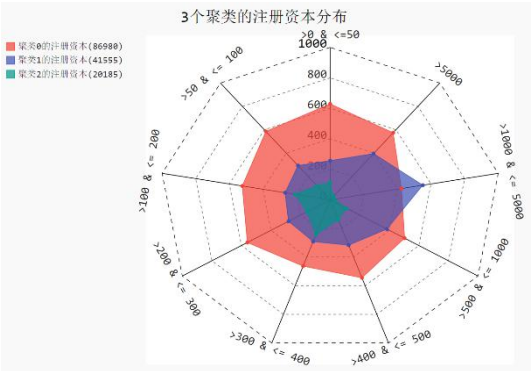

图 2-12　企业背景-注册资本

- 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 2-13~2-16 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，这是以为不同特征的量纲有时会差距过大。

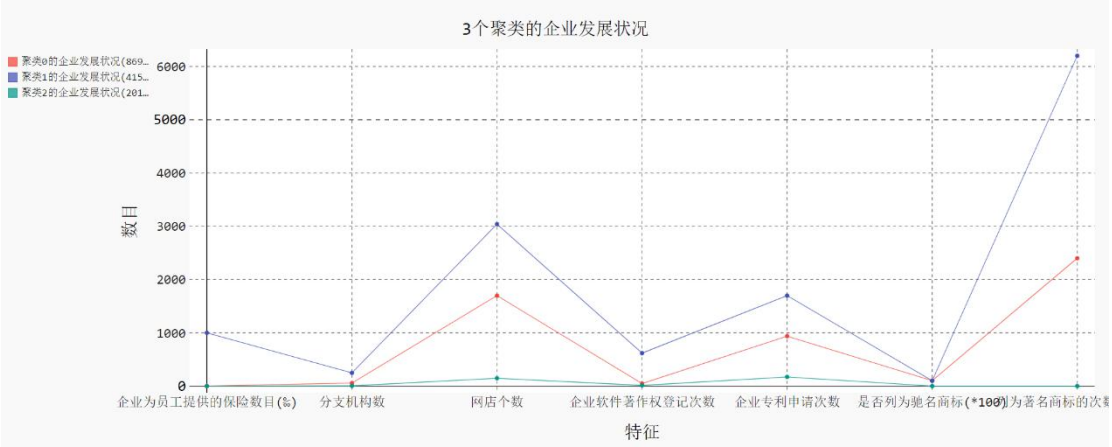

图 2-13　企业发展状况

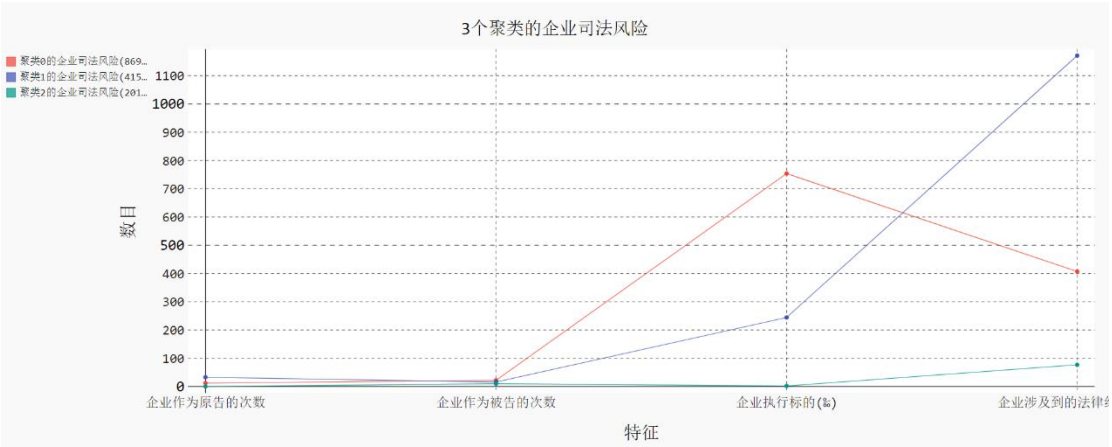

图 2-14　企业司法风险

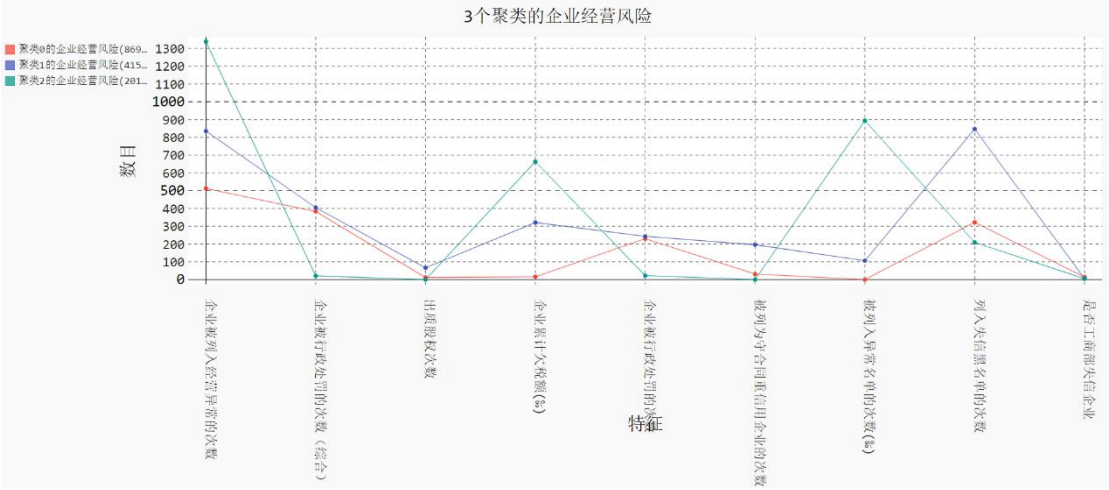

图 2-15　企业经营风险

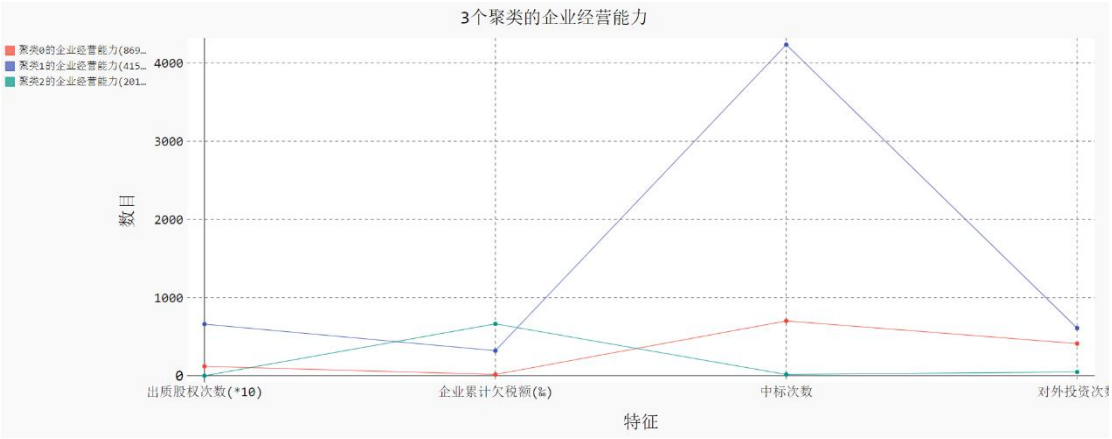

图 2-16　企业经营能力

图 2-17~2-20 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。

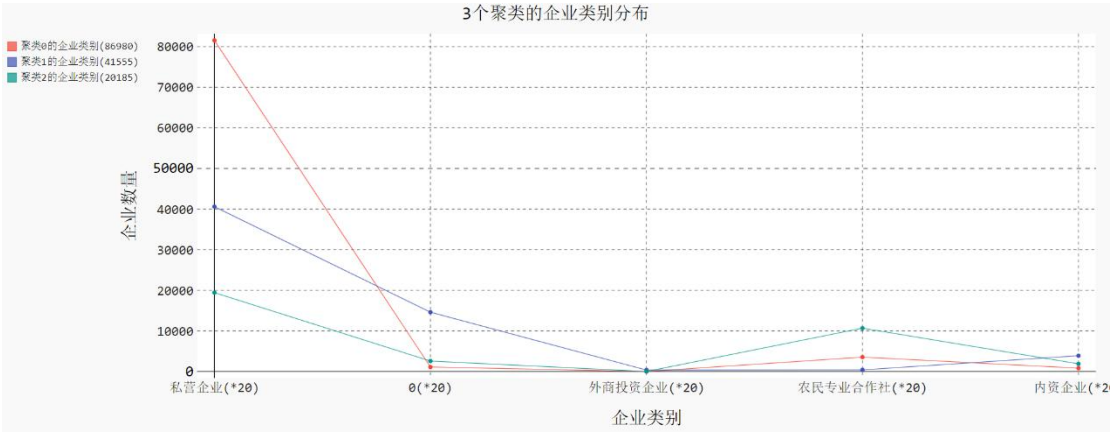

图 2-17　企业类别

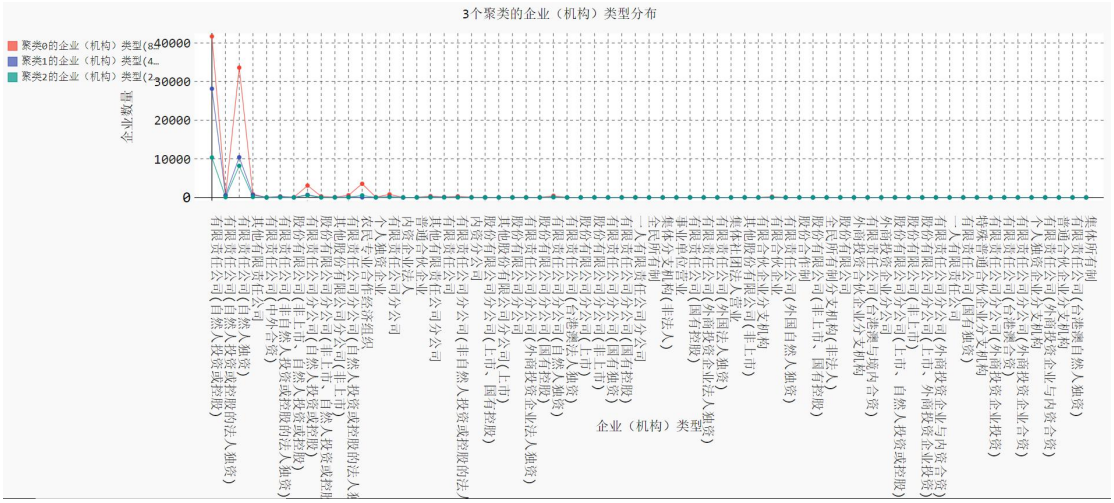

图 2-18　企业（机构）类型

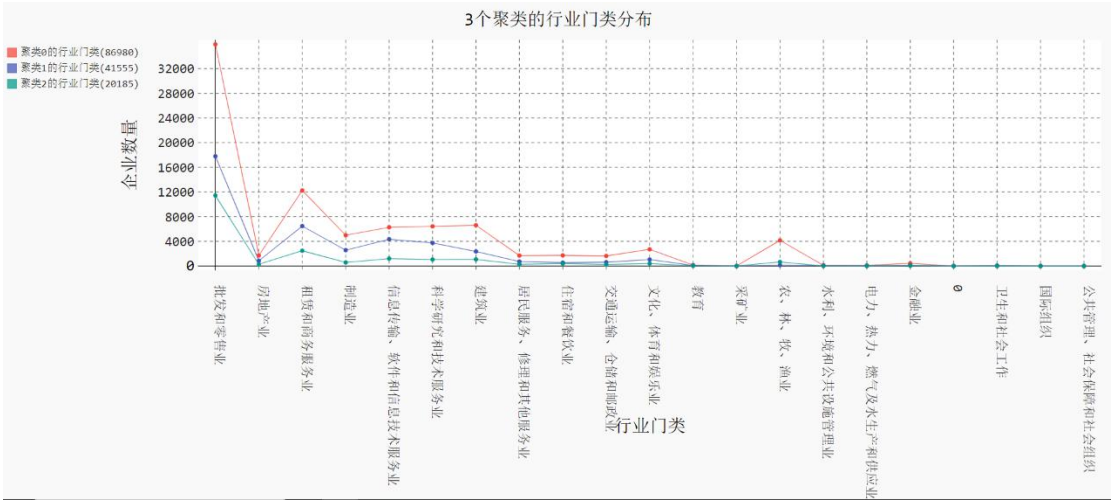

图 2-19　行业门类

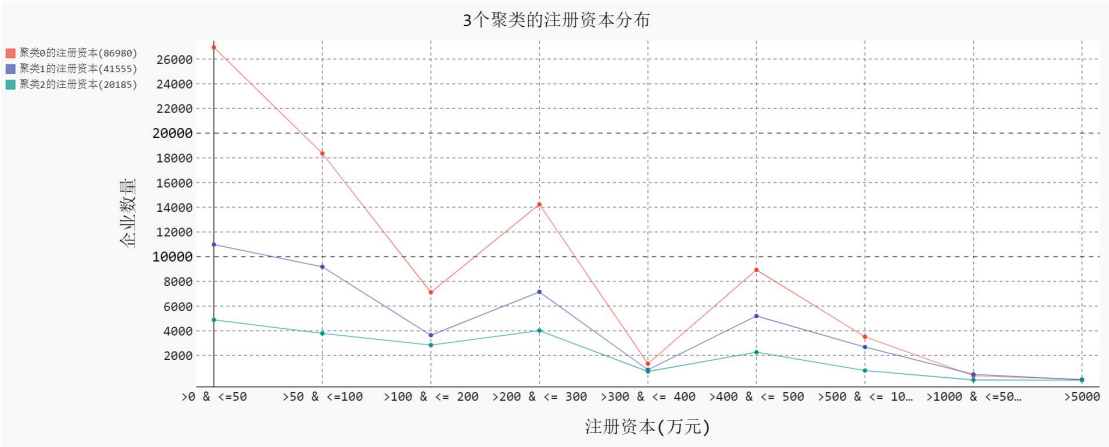

图 2-20　注册资本

- 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 2-21~2-25 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 2-26 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。

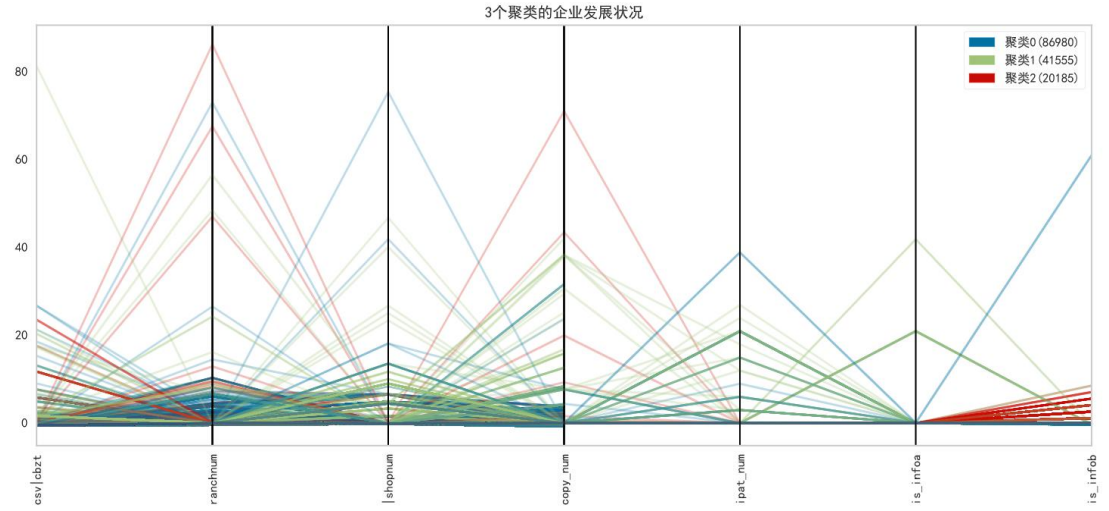

图 2-21　企业发展状况

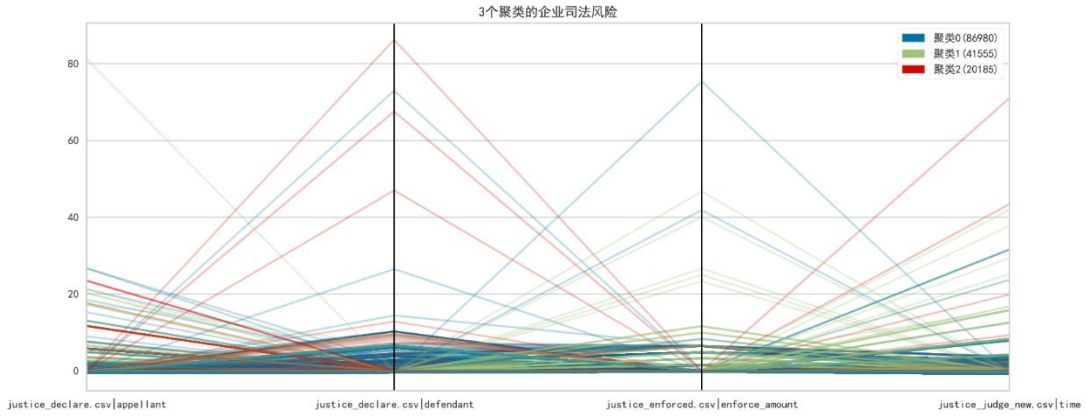

图 2-22 企业司法风险

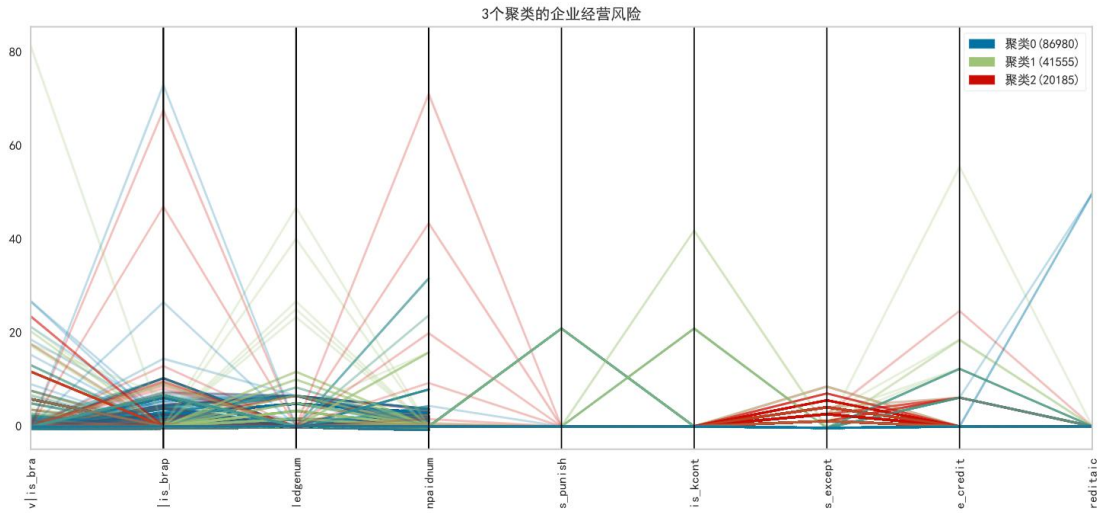

图 2-23 企业经营风险

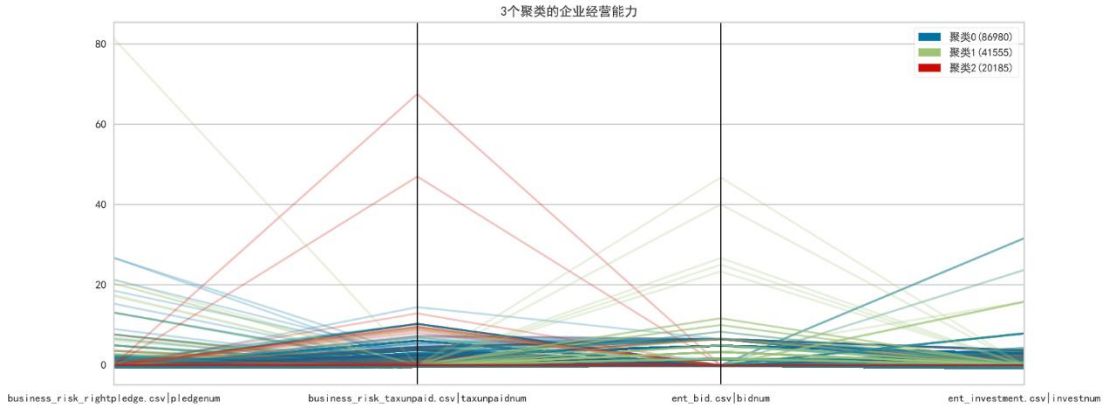

图 2-24 企业经营能力

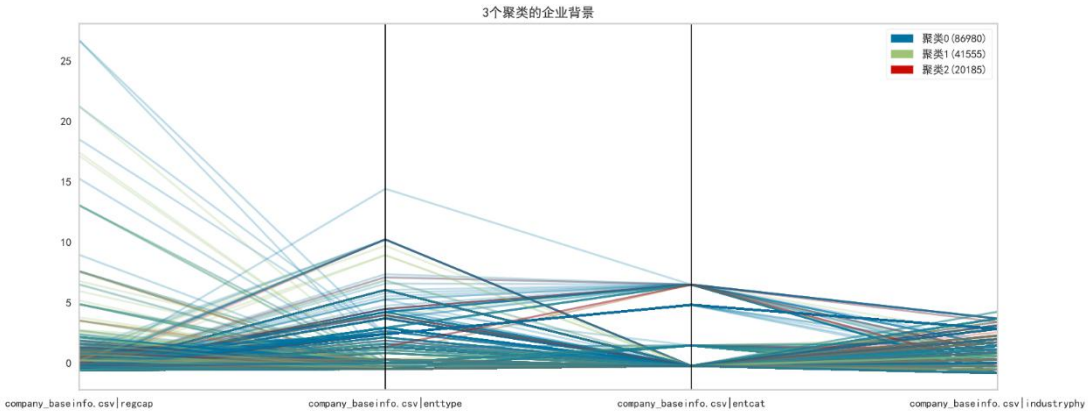

图 2-25　企业背景

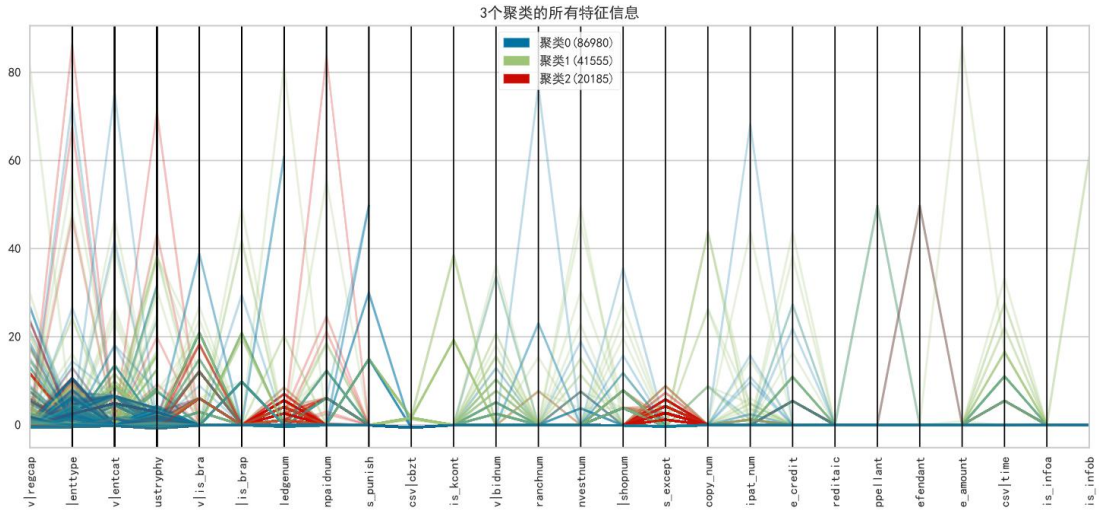

图 2-26　企业所有信息

针对上述数据分析，我们标注每类的标签，并通过交互词云可视化的形式呈现在 Web 端。以下分别为 0,1,2 类的标签结果：

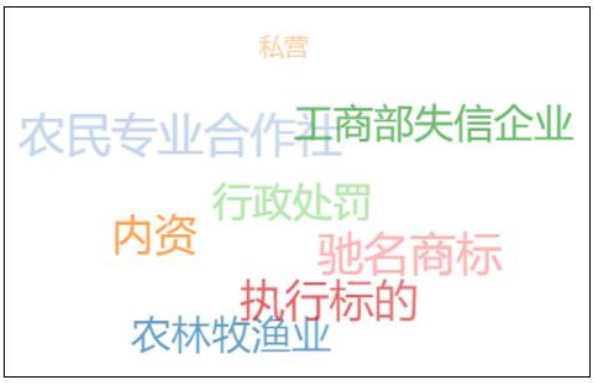

图 2-27　Class0 标签词云

图 2-28　Class1 标签词云

图 2-29　Class2 标签词云

3 Web 前端可视化

3.1 单个企业检索

检索单个企业，展示该企业的基本信息，通过颜色填充展示所在类别。

图 3-1 单个企业查询及结果展示

3.2 批量检索

根据部分特征检索批量企业，展示企业的基本信息及所在类别，可以通过按钮排序实现类别聚集展示。

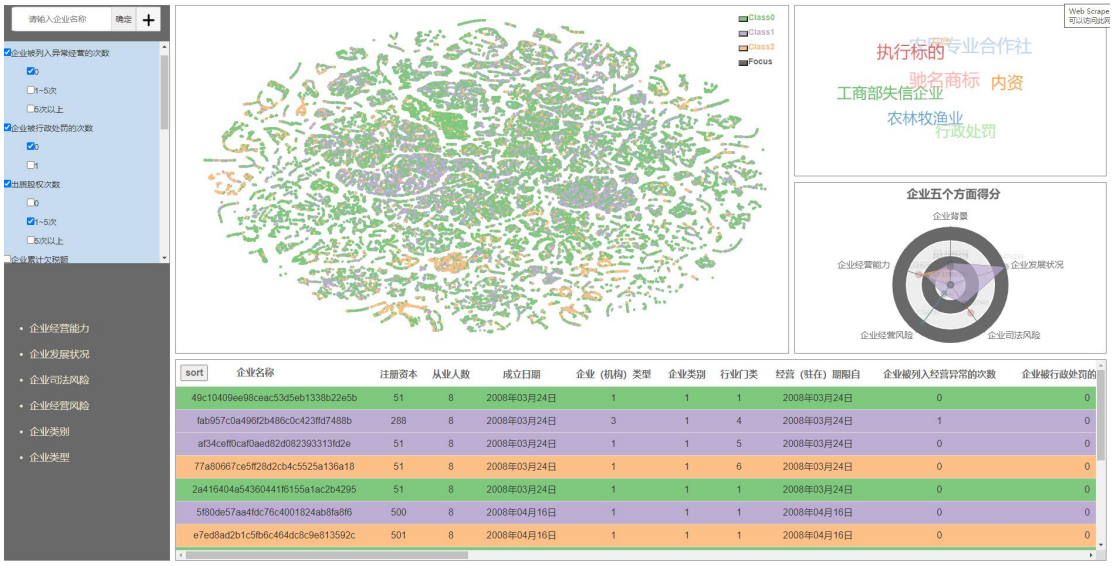

图 3-2 批量检索及结果展示

图 3-3 批量类别排序

3.3 TSNE 降维可视化

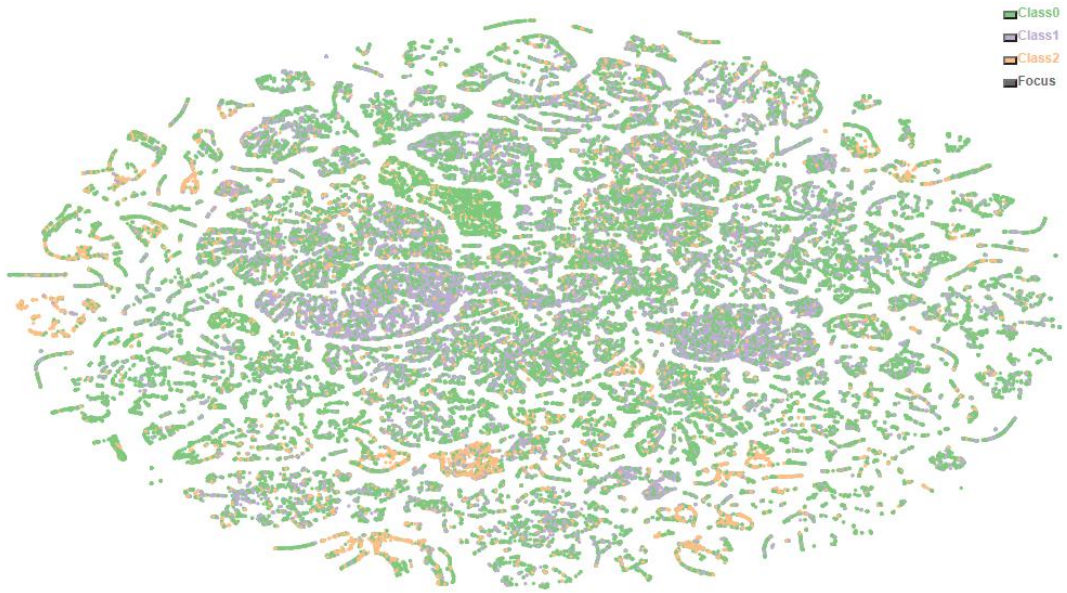

图 3-4 TSNE 降维可视化

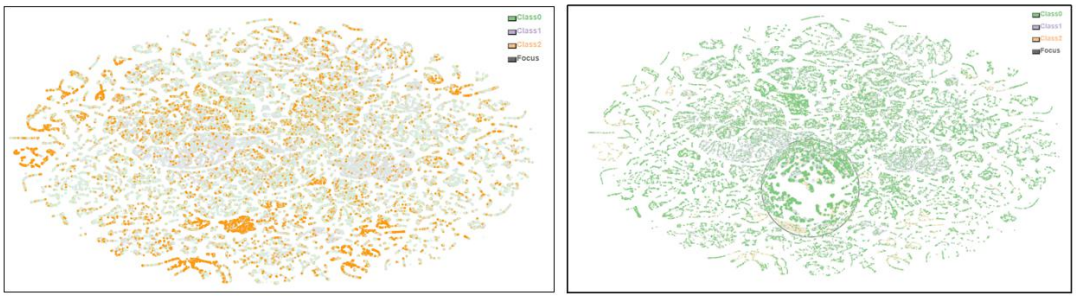

图 3-5 类别、鱼眼聚焦交互

3.4 标签词云可视化

图 3-6 标签词云交互

3.5 雷达图表分析

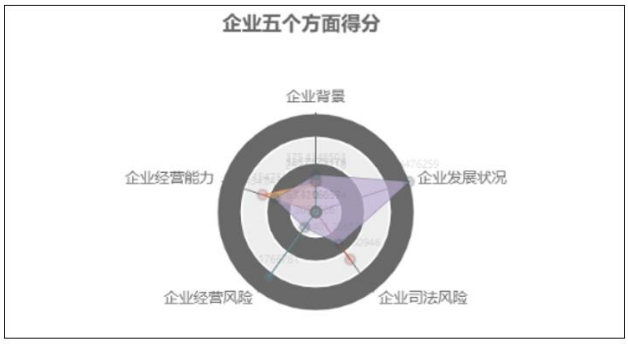

图 3-7 雷达图表分析