

**目录**

[一、引言 2](#_Toc40985456)

[二、基本原理 2](#_Toc40985457)

[2.1数据表分类 2](#_Toc40985458)

[2.2 K-Means聚类算法 5](#_Toc40985459)

[2.3 聚类性能评估 7](#_Toc40985460)

[2.3.1 Calinski-Harabasz（CH指数） 7](#_Toc40985461)

[2.3.2 Silhouette Coefficient（轮廓指数） 7](#_Toc40985462)

[2.3.3 Davies-Bouldin（DB指数） 8](#_Toc40985463)

[三、方法 9](#_Toc40985464)

[3.1等级化操作 9](#_Toc40985465)

[3.2模块统计 10](#_Toc40985466)

[3.3企业聚类 13](#_Toc40985467)

[3.3.1 等级聚类 13](#_Toc40985468)

[3.3.2模块属性聚类 13](#_Toc40985469)

[四、实验 13](#_Toc40985470)

[4.1实验数据与实验方法 13](#_Toc40985471)

[4.2实验结果与分析 13](#_Toc40985472)

[4.2.1企业总体聚类 13](#_Toc40985473)

[4.2.2企业总评聚类模型选择 14](#_Toc40985474)

[4.2.3标签分析 16](#_Toc40985475)

[4.2.4算法效率 24](#_Toc40985476)

[五、结论 28](#_Toc40985477)

[六、参考文献 28](#_Toc40985478)

**摘要：**金融场景中企业这一信贷主体的数据量大，来源广泛、涉及企业的维度丰富，在构建企业画像和分类的过程中遇到很大挑战。因此本文利用K-Means算法，针对企业画像构建和分类，提出一种加权多层K-Means技术来构建企业分类和画像构建。首先对企业数据进行分类，分成7个模块；再经过等级化预处理，统一模块属性度量，获得每个模块用于建模的指标；再将每个模块内的指标分别进行加权求和后再“等级化”处理，得到每个模块的等级；将模块等级加权求和获取企业总体水平，并对每个模块的等级矩阵通过加权K-Means聚类算法，实现簇划分。本文以出题方所给出的36张表作为数据源，处理后的结果表明训练模型在CH=136499.42、DB=0.87、SH=0.56，且训练时间为11.43(s)，预测时间为0.23(s),相较于传统的K-Means、Birch、GaussianMixture三种无监督模型，上述指标全方面领先，簇的分类区别明显。

**关键词：**企业数据，K-Means，等级化，无监督聚类

## 一、引言

金融场景中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，数据量大，来源广泛、涉及企业的维度丰富，在分析企业还款能力、信用水平过程中面临巨大的挑战。即在构建企业的画像的过程中，我们面临以下的问题：

（1）由于涉及的企业数据维度丰富，而真正能用于评估的，只是部分数据属性，一些属性还需要进行人工的预处理，才能提取出可以用数据建模分析的指标，加大了分析的难度。

（2）不同的企业在经营、风险等多个方面情况都有非常大的不同，导致不同的企业在不同的属性上，都可能存在一定的缺失情况，难以建立统一的标准，以指标的方式，对企业进行评估，同时也加大了利用企业属性的相似性对企业进行聚类。。

本赛题旨在寻找一种有效的无监督分类方法，能够

①建立统一度量的建模指标，便于企业标签的对建模指标贴标签；

②对企业属性根据内部特性进行模块分类，实现信息提炼，数据降维，加速聚类效率。

③实现加权聚类，使得算法人员能够根据每个属性的实际情况，调整其最终对聚类结果的影响。

基于此，本文设计一种加权多层K-Means算法，作为一种企业分类技术的理论支持。

## 二、基本原理

### 2.1数据表分类

针对以上问题，我们首先将出题方提供的数据表，分成7个模块，如表2-1-1至2-1-7所示：

表2-1-1：企业数据表归类（风险）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 模块内包含的表 | 解释 |
| 1-风险模块  (risk\_module) | administrative\_punishment | 行政处罚 |
| business\_risk\_abnormal | 企业经营风险-经营异常 |
| business\_risk\_all\_punish | 企业经营风险-行政处罚（综合） |
| business\_risk\_taxunpaid | 企业经营风险-欠税公告 |
| business\_risk\_rightpledge | 企业经营风险-股权出质 |
| ent\_social\_security | 年报社保信息（参保状态/年报五险一金欠税额） |
| exception\_list | 异常名单 |
| justice\_declare | 司法风险—开庭公告数据 |
| justice\_enforced | 司法风险—被执行人数据 |
| justice\_judge\_new | 司法风险-裁判文书数据 |
| justice\_credit | 司法风险-失信黑名单数据 |
| justice\_credit\_aic | 失信企业（工商部） |

表2-1-2：企业数据表归类（投资）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 模块内包含的表 | 解释 |
| 2-投资模块  (investment  \_module) | ent\_bid | 中标数据 |
| ent\_branch | 企业分支机构信息 |
| ent\_contribution | 企业出资信息(股东（自然人）出资信息) |
| ent\_contribution\_year | 企业年报出资信息 |
| ent\_guarantee | 企业年报对外担保 |
| ent\_investment | 年报对外投资 |
| ent\_onlineshop | 年报网店信息 |
| enterprise\_insurance | 单位参保信息查询（养老单位参保信息） |

表2-1-3：企业数据表归类（知识产权）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 模块内包含的表 | 解释 |
| 3-知识产权模块  (credit\_module) | intangible\_brand | 知识产权\_商标数据 |
| intangible\_copyright | 知识产权\_软件著作权数据 |
| intangible\_patent | 知识产权\_专利数据 |
| web\_record\_info | 知识产权\_域名数据 |

表2-1-4：企业数据表归类（品牌）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 模块内包含的表 | 解释 |
| 4-品牌模块  (brand\_module) | jn\_special\_new\_info | 济南市专精特新中小企业 |
| jn\_tech\_center | 济南市省级市级企业技术中心名录 |
| trademark\_infoa | 驰名商标信息 |
| trademark\_infob | 著名商标信息 |
| product\_checkinfo\_connect | 企业产品被抽查信息 |

表2-1-5：企业数据表归类（招聘）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 模块内包含的表 | 解释 |
| 5-招聘模块  (recruit\_module) | recruit\_zhyc | 招聘\_中华英才网 |
| recruit\_qcwy | 招聘\_前程无忧 |
| recruit\_zlzp | 招聘\_智联招聘 |

表2-1-6：企业数据表归类（信用）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 模块内包含的表 | 解释 |
| 6-信用模块  (credit\_module) | enterprise\_keep\_contract | 守合同重信用企业 |
| jn\_credit\_info | 济南市信用信息 |

表2-1-7：企业数据表归类（基本信息）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 模块内包含的表 | 解释 |
| 7-基本信息模块  (baseinfo\_module) | company\_baseinfo | 企业基本信息 |
| change\_info | 企业变更信息 |

通过模块分类的方式，将描述属性一致的表归入同一个模块，并筛选和预处理有价值的数据进行建模，评估某一个模块的水平。既能更加综合客观，也实现了维度降低，同时也极大程度上，降低了由于数据缺失带来的问题。

每张表具体的预处理方法，参照《预处理规则》第二章属性预处理。预处理后，每个模块中用于建模的属性如表2-2-1至2-2-7所示。

表2-2-1：模块建模属性（风险）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 评估属性 | 说明 |
| 1-风险模块  (risk\_module) | is\_punish | 行政处罚次数 |
| is\_bra | 经营异常次数 |
| pledgenum | 股权出质次数 |
| is\_brap | 行政处罚记录次数 |
| taxunpaidnum | 欠税总额 |
| unpaid\_sum | 欠缴保险总额 |
| is\_except | 异常次数 |
| declaredate | 最新纠纷日期 |
| appellant\_amount | 原告次数 |
| defendant\_amount | 被告次数 |
| enforce\_amount | 执行金额总额 |
| record\_date | 执行日期 |
| judge\_new\_count | 诉讼次数 |
| is\_justice\_credit | 工商部失信次数 |
| is\_justice\_creditaic | 司法风险失信次数 |

表2-2-2：模块建模属性（投资）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 评估属性 | 说明 |
| 2-投资模块  (investment  \_module) | insurance\_num | 年平均参保个数 |
| bidnum | 中标次数 |
| branchnum | 分支个数 |
| subconam\_total | 认缴总额 |
| liacconam | 累计实缴总额 |
| lisubconam | 累计认缴总额 |
| investnum | 对外投资次数 |
| shopnum | 网店个数 |

表2-2-3：模块建模属性（知识产权）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 评估属性 | 说明 |
| 3-知识产权模块  (creativity\_module) | ibrand\_num | 商标申请次数 |
| icopy\_num | 软著登记次数 |
| ipat\_num | 专利申请次数 |
| idom | 域名知识产权个数 |

表2-2-4：模块建模属性（品牌）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 评估属性 | 说明 |
| 4-品牌模块  (brand\_module) | is\_jnsn | 济南市中专精小企业 |
| level\_rank | 科技等级 |
| is\_infoa | 是否驰名商标 |
| is\_infob | 是否著名商标 |
| passpercent | 质检通过率 |

表2-2-5：模块建模属性（招聘）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 评估属性 | 说明 |
| 5-招聘模块  (recruit\_module) | qcwynum | 前程无忧的招聘记录数 |
| zhycnum | 中华英才的招聘记录数 |
| zlzpnum | 智联招聘的招聘记录数 |

表2-2-6：模块建模属性（信用）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 评估属性 | 说明 |
| 6-信用模块  (credit\_module) | is\_kcont | 是否守合同重信用企业 |
| credit\_grade | 信用等级 |

表2-2-7：模块建模属性（基本信息）

|  |  |  |
| --- | --- | --- |
| 模块名称 | 评估属性 | 说明 |
| 7-基本信息  (company\_baseinfo  \_module) | regcap | 注册资本 |
| empnum | 从业人数 |
| esdate | 成立日期 |

### 2.2 K-Means聚类算法

在本系统中，企业的聚类算法，在经过实验对比后，最终确定K-Means。

K-Means算法是一种无监督学习，同时也是基于划分的聚类算法，一般用欧式距离作为衡量数据对象间相似度的指标，相似度与数据对象间的距离成反比，相似度越大，距离越小。算法需要预先指定初始聚类数目，根据数据对象与聚类中心之间的相似度，不断更新聚类中心的位置，不断降低类簇的误差平方和（Sum of Squared Error, SSE），当SSE不再变化或目标函数收敛时，聚类结束，得到最终结果。

K-Means算法的核心思想是：首先从数据集中随机选取个初始聚类中心()，计算其余数据对象与聚类中心的欧氏距离，找出离目标数据对象最近的聚类中心，并将数据对象分配到聚类中心，所对应的簇中。然后计算每个簇中数据对象的平均值作为新的聚类中心，进行下一次迭代，直到聚类中心不再变化或达到最大的迭代次数停止。

空间中数据对象与聚类中心间的欧式距离计算公式为：

 （2-1）

式（2-1）中，为数据对象，为第个聚类中心，为数据对象的维度，,为和的第个属性值。

整个数据集的误差平方和计算公式为：

 （2-2）

式（2-2）中，的大小表示聚类结果的好坏，k为簇的个数。当SSE的值不再变化，或者变化小于一定阈值时，停止迭代。其流程如图2-1所示。

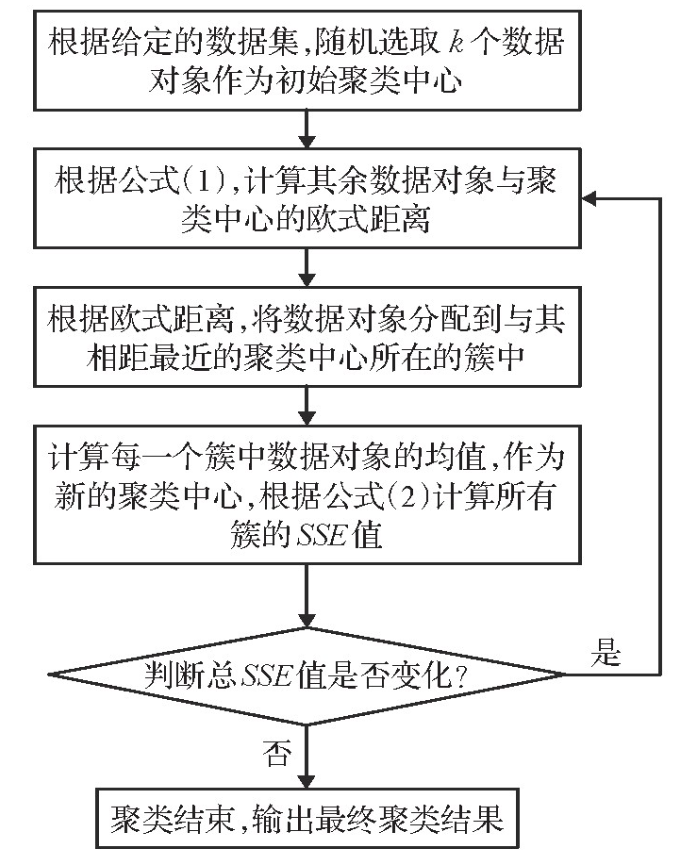


图2-1：K-Means聚类算法流程

### 2.3 聚类性能评估

#### 2.3.1 Calinski-Harabasz（CH指数）

类别内部数据的协方差越小越好，类别之间的协方差越大越好（换句话说：类别内部数据的距离平方和越小越好，类别之间的距离平方和越大越好），

这样的Calinski-Harabasz分数s会高，分数s高则聚类效果越好，s的求解公式如（2-3）所示：

 （2-3）

式（2-3）中表示簇的个数为的分数。是矩阵的迹，迹的定义如式（2-4）和（2-5）所示，是为类别之间的协方差矩阵，为类别内部数据的协方差矩阵，为训练集样本数，为类别数

 （2-4）

 （2-5）

#### 2.3.2 Silhouette Coefficient（轮廓指数）

轮廓系数（Silhouette Coefficient），是聚类效果好坏的一种评价方式。最早由 Peter J. Rousseeuw 在 1986 提出。它结合内聚度和分离度两种因素。可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。

1) 计算样本到同簇其他样本的平均距离。越小，说明样本越应该被聚类到该簇。将称为样本的簇内不相似度。簇中所有样本的均值称为簇的簇不相似度。

2) 计算样本到其他某簇 的所有样本的平均距离，称为样本与簇的不相似度。定义为样本的簇间不相似度为式(2-6)所示：

 （2-6）

式(2-6)中表示个簇。越大，说明样本越不属于其他簇。

3) 根据样本的簇内不相似度和簇间不相似度，定义样本的轮廓系数为式(2-7)：

 （2-7）

接近1，则说明样本聚类合理；接近-1，则说明样本更应该分类到另外的簇；若近似为0，则说明样本在两个簇的边界上。

#### 2.3.3 Davies-Bouldin（DB指数）

DB(Davies-Bouldin)由Davies 和Bouldin提出，是一种基于类内相似性和类间差异性的有效性评价。对于聚类划分，为其第个类别划分，,为总类别数。设为的分散程度，为类别和的不相似程度，定义和间的相似性指标满足下述5个条件。

1. ；
2. ；
3. 若且，；
4. 若且，；
5. 若且，；

条件1）、2）表明非负且对称，条件3）表明，当2类别均重叠于一点时，，条件4）表明，与其他2类别和差异性相同的类别，与具有较大分散度的类别更相似，条件5）表明，两分散度相同的类别和，类别更相似于差异性较小的类别。令

 （2-8）

 （2-9）

则一种简单的可表示为（2-10）所示的公式：

 （2-10）

那么的最大值可以表示为式（2-11）所示：

 （2-11）

则表示为式（2-12）所示的公式：

 （2-12）

上述各式中，为类别的类别中心点，为类别的数据个数。较小的意味着聚类结果中各类别具有较好的类内相似性和类间差异性。

## 三、方法

### 3.1等级化操作

在加权多层K-Means算法中，等级化操作，利用K-Means算法对企业一个维度的属性进行聚类，并按照结果进行排序，实现企业属性量纲的统一，体现了加权多层K-Means中的“K-Means”。

由于在企业画像构建过程中涉及到多个属性和中间属性，每个属性量纲不同，所以需要先进行等级化处理，将企业数据按照一定排列方式，分成多个等级。定义以下等级化操作，对于某一个企业属性，拥有条企业的记录，可以定义如下集合如式（3-1）：

 （3-1）

式（3-1）中是个企业在该属性的集合，是第个企业在该属性上的值。

使用K-Means算法进行聚类，分成*n\_cluster*个类。再根据每个簇的中心的值，升序排列，即根据簇中心的高低，赋予1到*n\_cluster*的值，将这种操作记为*sorted\_K-Means*，即划分*n\_cluster* = 5个等级，对模块总评和企业最终总评划分*n\_cluster* = 10个等级。对于原来是空值或者0的记录，赋予0，将该处理函数记作，如式（3-2）所示：

 （3-2）

记函数处理后的属性序列为新等级化序列为，如式（3-3）所示：

 （3-3）

式（3-3）中表示第个企业属性的等级。

### 3.2模块统计

在加权多层K-Means算法中，模块统计部分体现了“加权多层”的思想。

根据表2-2-1到2-2-7的属性分类，将不同的属性归入到个不同的模块当中，令模块集合为，共有个企业，则表示为式（3-4）所示：

 （3-4）

式（3-4）中表示第个企业的所有模块，表第个企业，在模块集合中的第个模块。对于，用表示模块内的属性集合，设该模块内有个属性，则可以表示为式（3-5）：

 （3-5）

式（3-5）中表示中第个属性的值，对应图3-1中的原属性层。将其经过等级化后的属性集合可以表示为式（3-6）：

 （3-6）

式（3-6）中表示模块中第个属性的值，即原属性经过等级化操作后，图3-1中的原属性等级层。即“加权多层”中的第一层。

由于每个属性对最终属性评估的影响程度不同，因此，对每个属性赋予不同的权重，以模块为例，拥有个属性，则权重集合如式（3-7）所示：

 （3-7）

式（3-7）中表示模块第个属性的权重。则模块的加权总分如式（3-8）所示：

 （3-8）

式（3-8）中即第个企业在第个模块上的加权总分，对应图3-1中的模块评分层，此处对应“加权多层”中的第一次加权。

由于每个模块属性的个数不同，导致最后不同的模块仍然存在量纲不统一的问题，此时再对模块总分，进行式（3-3）所示的函数操作，赋予其*n\_cluster* = 10个等级，得到新的模块等级表示为式（3-9）所示：

 （3-9）

式（3-9）中对应图3-1中的模块等级层，对应“加权多层”中的第二层。

获得了每个模块的评分等级，就可以去求企业所有模块的总评分数。考虑到不同的模块可能对企业总评的影响程度不一致，所以需要对每个模块赋予不同的权重，调整其对企业总评的影响，模块权重如式（3-10）所示：

 （3-10）

式（3-10）中表示第个模块等级的权重。企业的所有模块加权总分如式（3-11）所示：

 （3-11）

式（3-11）中对应图3-1中的企业总评层，即企业总分，在求总分的过程对应“加权多层”中的第二次加权。

考虑到每个企业的最后的加权总分，不像满分100分一样能直观的感受企业在行业中的总体水平，因此，对最终获得的企业加权总分执行式（3-3）中的函数操作，赋予*n\_cluster* = 10个等级，记作，即企业总评等级，方便更直观的展示企业水平，如式（3-12）所示：

 （3-12）

式（3-12）对应图3-1中的企业等级层，对应“加权多层”中的第三层。

模块统计数据流见图3-1。

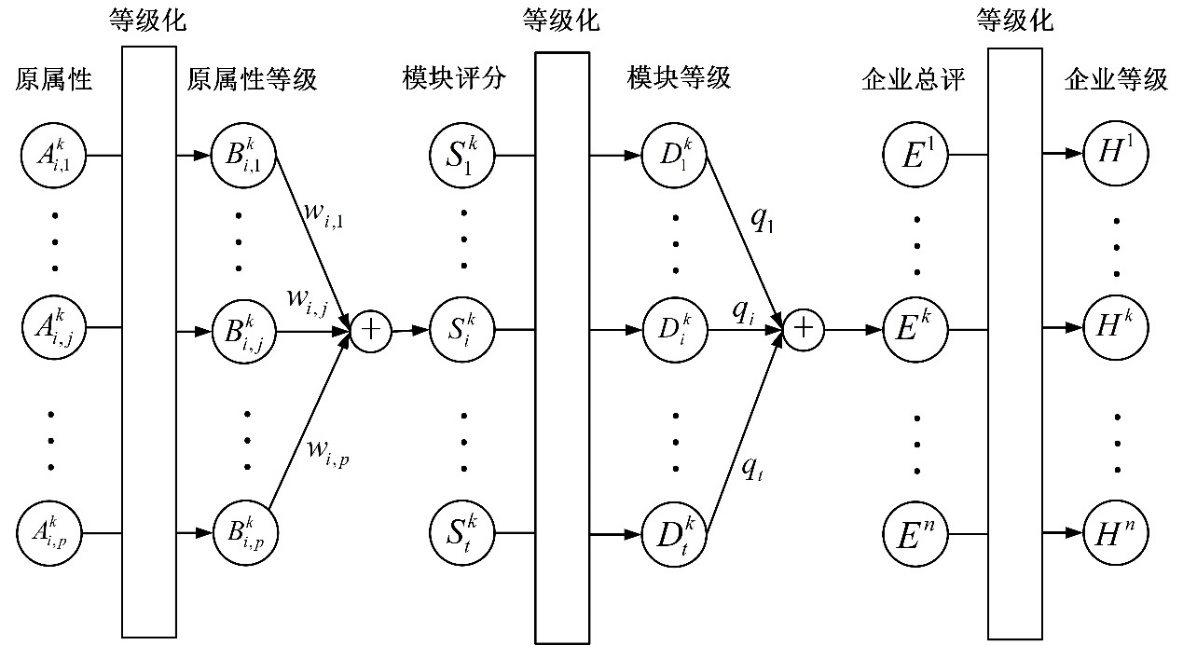


图3-1：算法数据流

模块统计部分的流程图如图3-2所示。

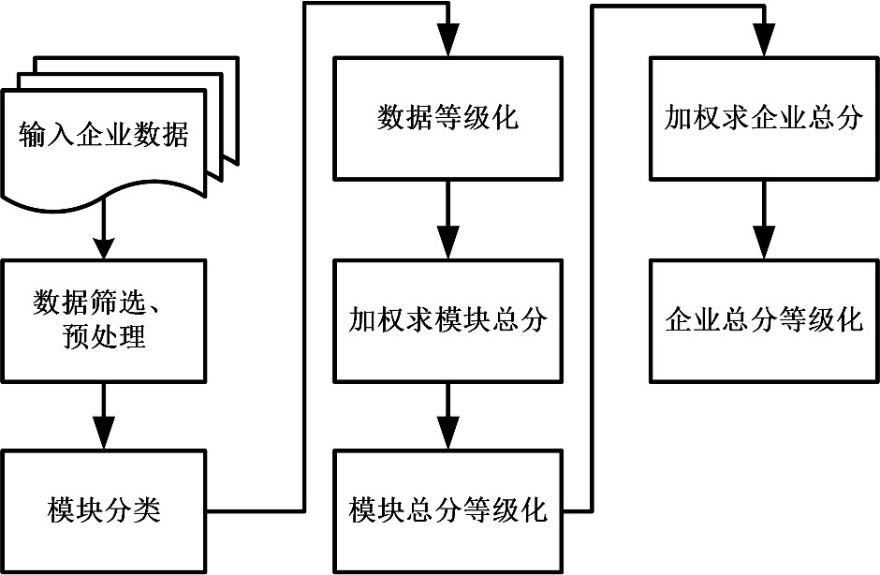


图3-2：算法流程图

### 3.3企业聚类

#### 3.3.1 等级聚类

显然，模块评分在经过等级化后得到了K个等级，则每个等级可以作为模块属性相似的依据；企业总评也在经过等级化后，也可以作为企业相似的依据。

#### 3.3.2模块属性聚类

由于每个企业拥有7个模块，可以采用等级化模块属性作为评估依据，结合属性权重，使用K-Means算法，对这个7维的数据，即7×n的矩阵进行加权聚类，将得到相同标签的企业分成一个簇，表示企业整体较为相似。可以用这个结果，作为企业推荐的依据。此处也体现了加权多层K-Means中的“加权K-Means”的思想。

## 四、实验

### 4.1实验数据与实验方法

本实验数据采用出题方所提供的36张表内共189038家企业的数据，内容涵盖了企业背景、商业贸易、知识产权、行政风险、司法风险、科技、招聘等多个方面的综合信息。

每张表的数据预处理方式，见《预处理方法》第二章。处理后的建模属性，见表2-2。

在企业聚类模型的研究中，由于等级聚类仅涉及一维，实际上类似给企业的属性打5个等级（模块等级和企业总评10个等级），是约定的规则，所以下面的实验，主要围绕模块属性聚类的企业总体聚类和模块聚类展开。

在模型训练数据选取上，我们采用交叉验证法，例如10折交叉验证(10-fold cross validation)，将数据集分成十份，轮流将其中9份做训练1份做验证，10次的结果的均值作为对算法精度的估计，一般还需要进行多次10折交叉验证求均值，例如：10次10折交叉验证，以求更精确一点。

聚类模型的选择上，选择K-Means，Birch，GaussianMixture三种聚类方法。其中K-Means是一种迭代求解的聚类方法；Birch是一种综合的层次聚类算法，GaussianMixture是一种高斯混合模型。三种聚类模型原理不同，从不同的角度进行聚类，因此选择这三种模型进行对比。通过CH（Calinski-Harabasz）指数，SH（Silhouette Coefficient）指数，DB（Davies-Bouldin）指数，综合评估性能，选择出合适的聚类方法，最后通过进行参数优化，构建合适的模型。

### 4.2实验结果与分析

#### 4.2.1企业总体聚类

企业总体评估的属性及权重如表4-1所示，表4-1中，用xxx\_module\_type，表示模块被等级化处理后的评级，其中由于风险模块对企业的评价起负作用，因此设置权重为负值，同时为了避免最后的加权求和得到的结果为0，与空值混淆，所以设置权重为-1.11。此外7个模块都是默认设置1-10个等级，企业等级由于组成复杂，为了区分度高设置10个等级，两个等级设置均不包含0或空值的等级0。

得到的ent\_inner\_type，就是通过对7个模块的属性进行聚类，得到的标签。

表4-1：企业总评属性表

|  |  |  |  |
| --- | --- | --- | --- |
| 表名 | 属性名 | 备注 | 权重 |
| ent | risk\_module\_type | 风险等级 | -1.11 |
| investment\_module\_type | 投资等级 | 1 |
| creatity\_module\_type | 知识产权等级 | 1 |
| brand\_module\_type | 品牌等级 | 1 |
| recruit\_module\_type | 招聘等级 | 1 |
| credit\_module\_type | 信用等级 | 1 |
| base\_info\_type | 基本信息等级 | 1 |
| ent | 企业总分 | / |
| ent\_type | 企业等级（10个等级） | / |
| ent\_inner\_type | 企业聚类标签 | / |

#### 4.2.2企业总评聚类模型选择

分别使用K-Means，Birch，GaussianMixture对获得的7个维度的企业数据集进行聚类。由于这些模型需要事先设置簇的个数，为了确定最佳的簇个数，我们预先设置了n\_cluster=[6,15]的整数集合进行尝试，对7×n（7个维度，n条记录）的矩阵进行聚类，并使用CH（Calinski-Harabasz）指数，SH（Silhouette Coefficient）指数（轮廓系数），DB（Davies-Bouldin）指数进行评估。

其中CH、SH值越大表示聚类效果越好，DB值越小表示聚类效果越好。如图4-1的(a)、(b)、(c)分别对应CH、DB、SH在簇变化时候，三种不同模型评估值的变化情况。

对于K-Means模型，最终簇的个数选择，综合考虑CH、DB、SH。由于三者值域并不同，因此对三者分别做归一化处理，得到CH’，DB’，SH’。根据单调性，CH、SH越大越优，DB越小越优，因此自定义指数CI如公式4-1所示：

 （4-1）

该自定义指标可以综合考虑三种指标，选中聚类模型最佳的簇个数。CI指标对应图4-1(d)。

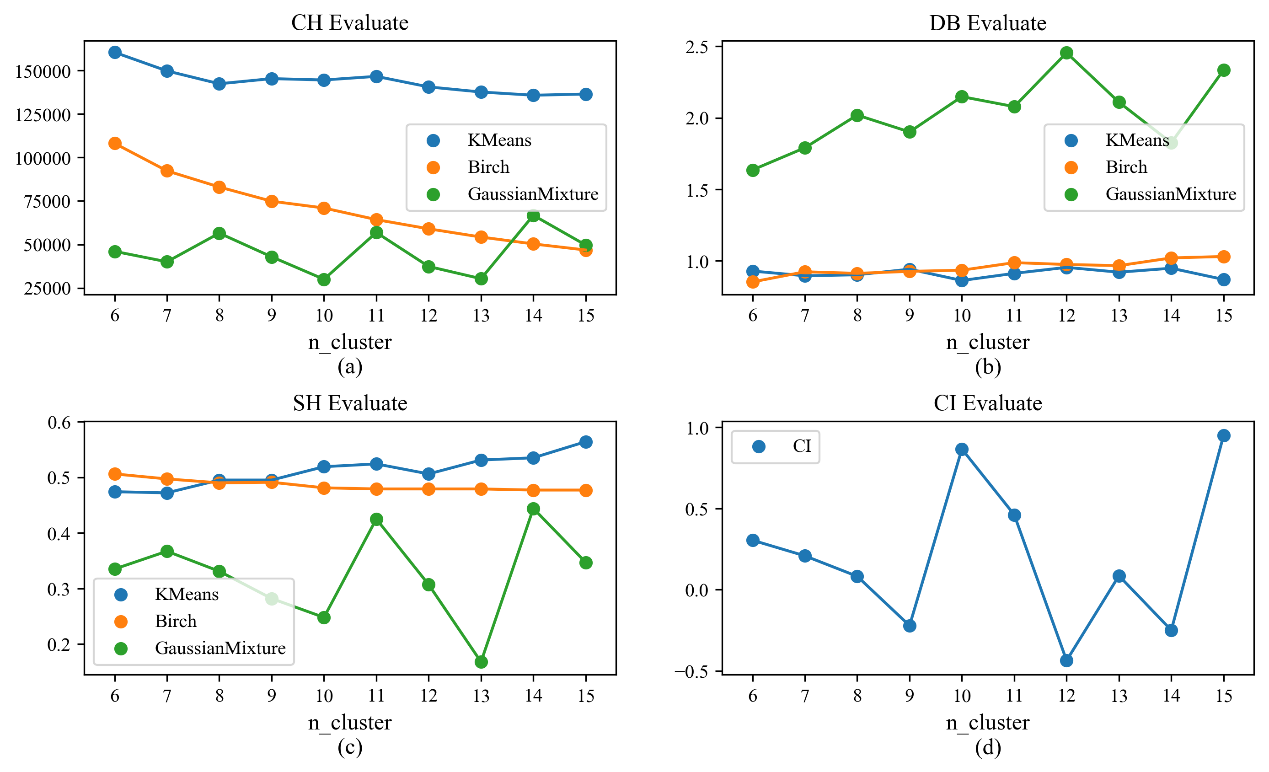


图4-1：三种聚类模型在不同簇值下CH、DB、SH的分布曲线. (a)三种聚类模型CH与簇个数关系性能的评估；（b）三种聚类模型DB与簇个数关系性能的评估；（c）三种模型SH与簇个数关系性能的评估；（d）K-Means在自定义指数CI与簇个数关系的性能评估

显然的，根据图4-1的(a)、(b)、(c)，在CH和SH中K-Means曲线基本在另外两个模型之上，而DB中的变化曲线在另外两个模型之下。因此，K-Means聚类算法效果优于Birch和GaussianMixture，于是选择K-Means模型继续分析。

显然，由CH’，DB’，SH’的单调性可知，CI的值与聚类效果好坏呈单调递增关系。K-Means的CI值随簇个数的变化关系，如图4-1(d)所示，综合考虑到簇的个数与CI的值，选择n\_cluster=15作为企业总评聚类的簇个数。此时CH= 136499.419；DB= 0.868；SH= 0.564。

PCA降维，并将降维的两个维度数据归一化后，结果如图4-2所示：

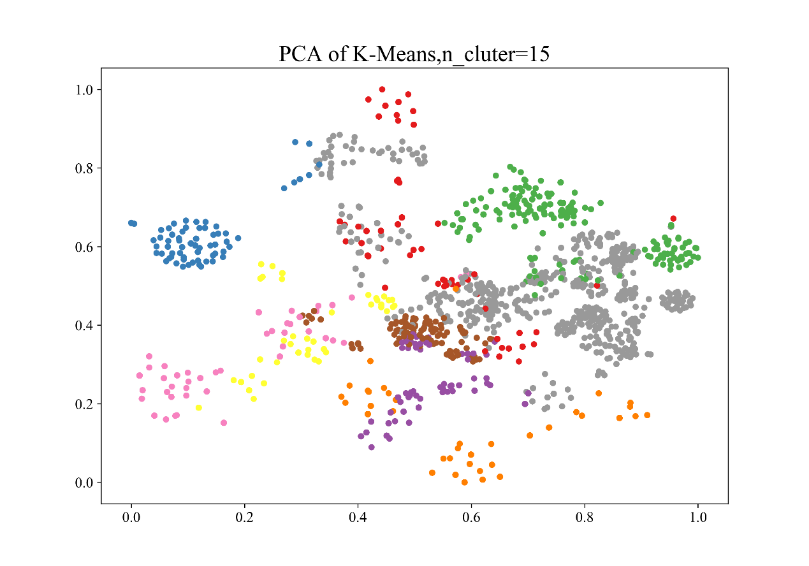


图4-2：PCA降维企业7个模块结果

图4-2中，不同颜色的点代表不同的簇，共有15个簇。可以看出整体每个类的边界较为明显，聚类效果较好。

K-Means模型的其他参数如下：

* max\_iter：最大迭代次数，设置为300次。
* tol：距离收敛值，使用默认的1×10-4。
* 距离计算公式：欧氏距离

#### 4.2.3标签分析

确定了聚类簇的个数后，需要分析每个簇中企业的数据含义，在这一部分中，将会对企业数据分布和每个簇中企业的数据分布进行分析。

考虑到企业总评估数据量庞大，且数据并不像子属性一样缺失的一样厉害，所以在模块评估和企业总分等级的划分上，划分成10个等级，更加合理。

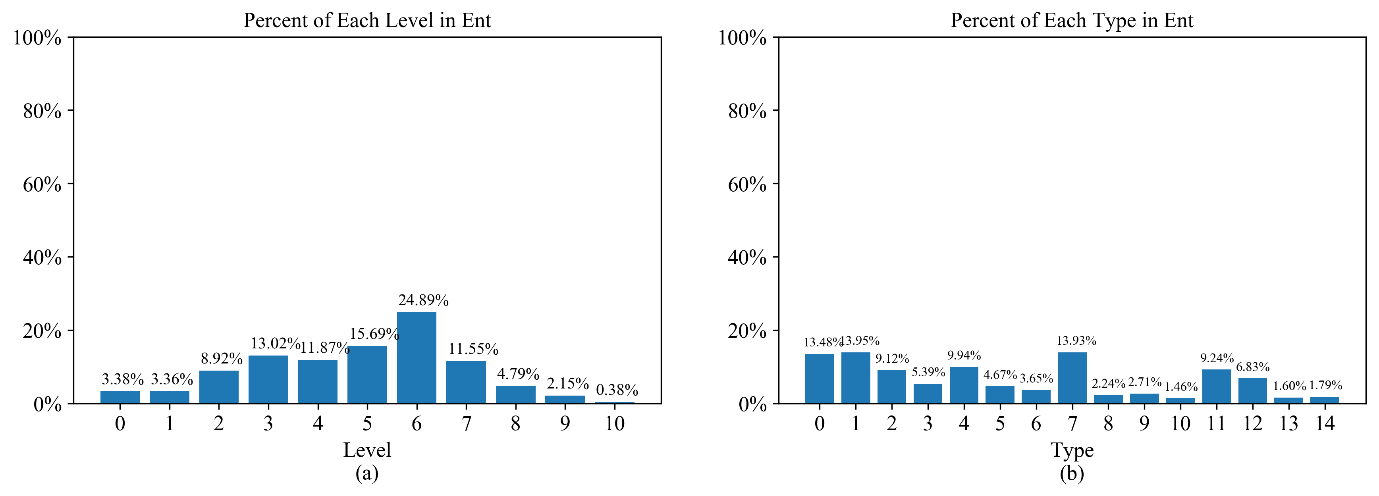


图4-3：企业等级分布与簇分布. (a)等级0-10的企业在总体中的分布；(b)15个簇的企业在总体中的分布

图4-3(a)表示在分出的10个等级中，每个等级的企业百分比分布情况，等级在3-7占大部分，反映了企业的平均水准。

图4-3(b)表示在分出的15个簇中，每个簇企业所占的百分比。

在显示15个簇的情况时，由于维数高于三维，且常用的降维手段（例如PCA）在数据可视化方面降维的数值并没有具体的含义，为了更好的了解每个簇内企业数据的分布，接下来在显示每个簇的特征时，采用柱状图的方式，呈现每个簇企业的个数，簇中企业在每个模块中等级的分布，以及该簇中企业总评等级的分布。

每个簇，都用一张图来描述，每张图都有8个子图，(a)对应风险等级；(b)对应投资等级；(c)对应知识产权等级；(d)对应品牌等级；(e)对应招聘等级，也可以理解为用人需求；(f)信用等级；(g)对应基本信息(规模)等级，也可以理解为企业资产等级；(h)对应企业总体等级。横坐标是指对应的等级，等级0表示记录为空，纵坐标表示该等级在该簇中所占的百分比。其中风险模块是等级越高，表示风险越大。其中最后一幅图(h)用红色标出，表示的是企业总体水平在该簇中的分布。

用这种方式，就可以知道每个簇，在每个模块下，等级分布情况，有助于我们更好的提炼簇的标签特征。

* 0号簇：共25477条记录。

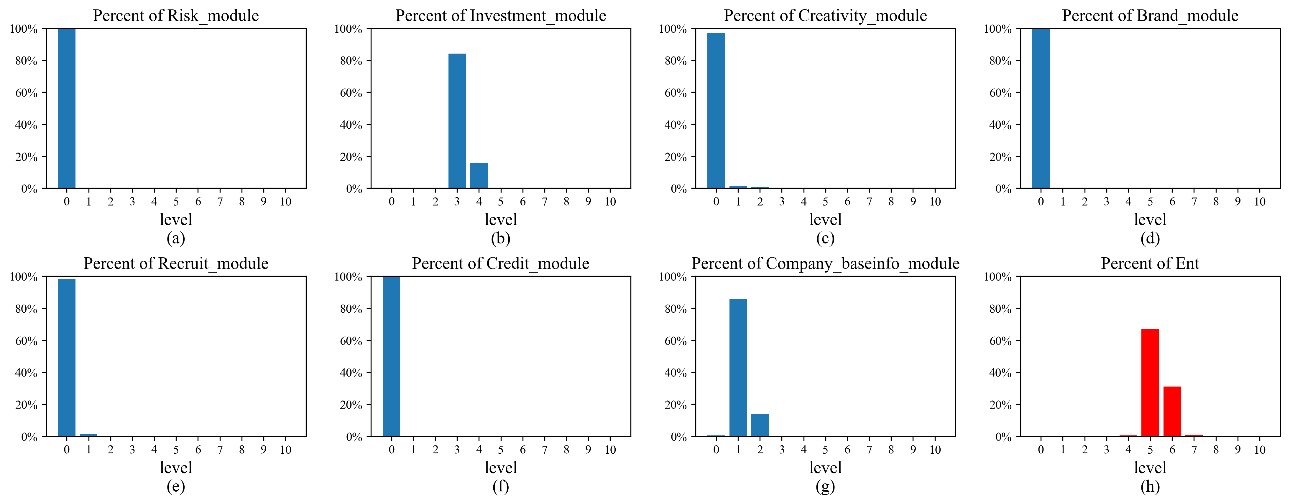


图4-4-1：企业总评0号簇各属性分布；(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

如图4-4-1所示，0号簇标签：中等偏低投资水平，资产等级水平较低；总体企业水平中等（图h）。

* 1号簇：共26362条记录。

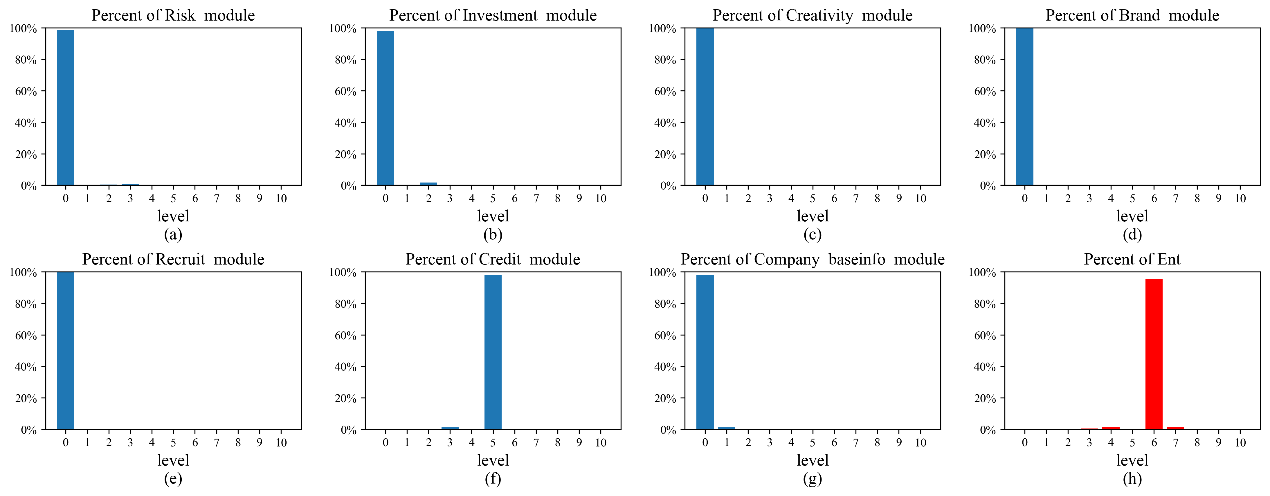


图4-4-2：企业总评1号簇各属性分布；(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

如图4-4-2所示，1号簇标签：中等信用等级；总体企业水平中等。

* 2号簇：共17246条记录。

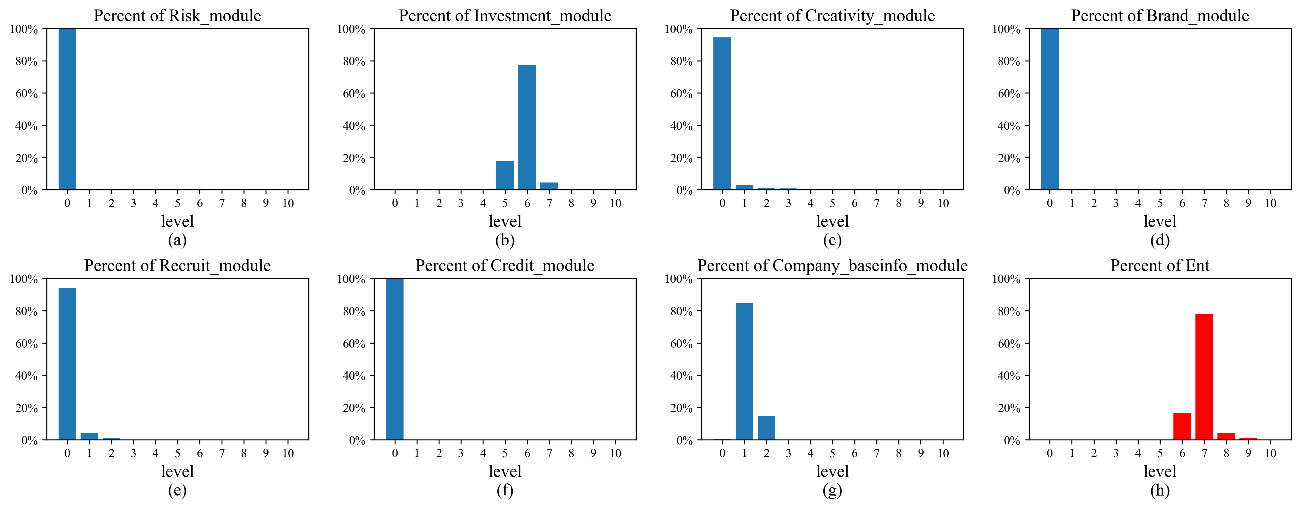


图4-4-3：企业总评2号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

如图4-4-3所示，2号簇标签：中等偏高的投资水平，较低的创新水平，有用人需求，较低的资产等级；总体企业水平中等偏高。

* 3号簇：共10188条记录。

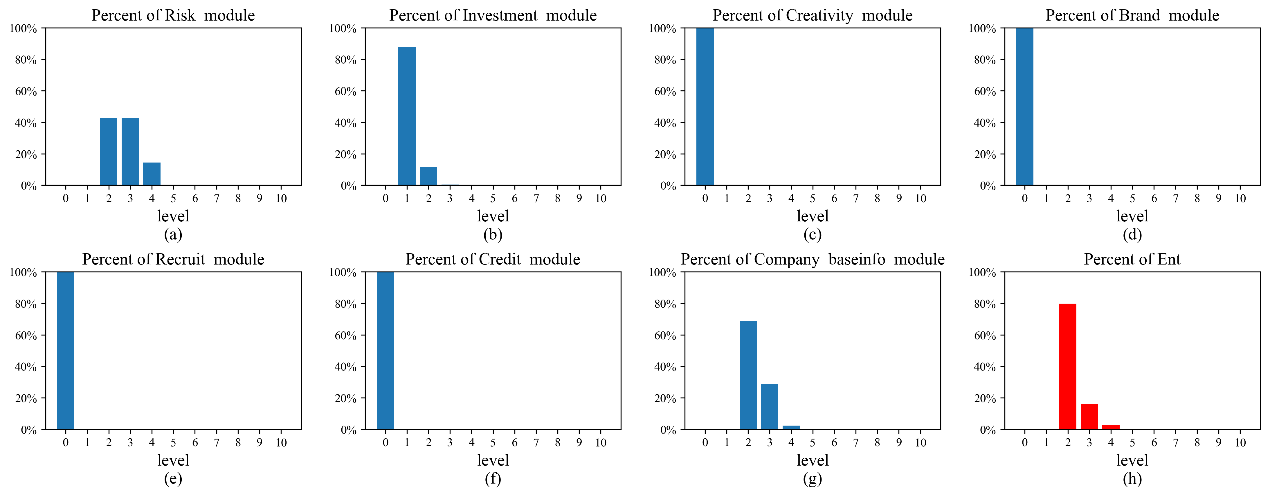


图4-4-4：企业总评3号簇各属性分布；(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

如图4-4-4所示，3号簇标签：中等偏低的风险水平，中等偏低的投资水平，较低的资产等级；企业总体水平中等偏低。

* 4号簇：共18793条记录。

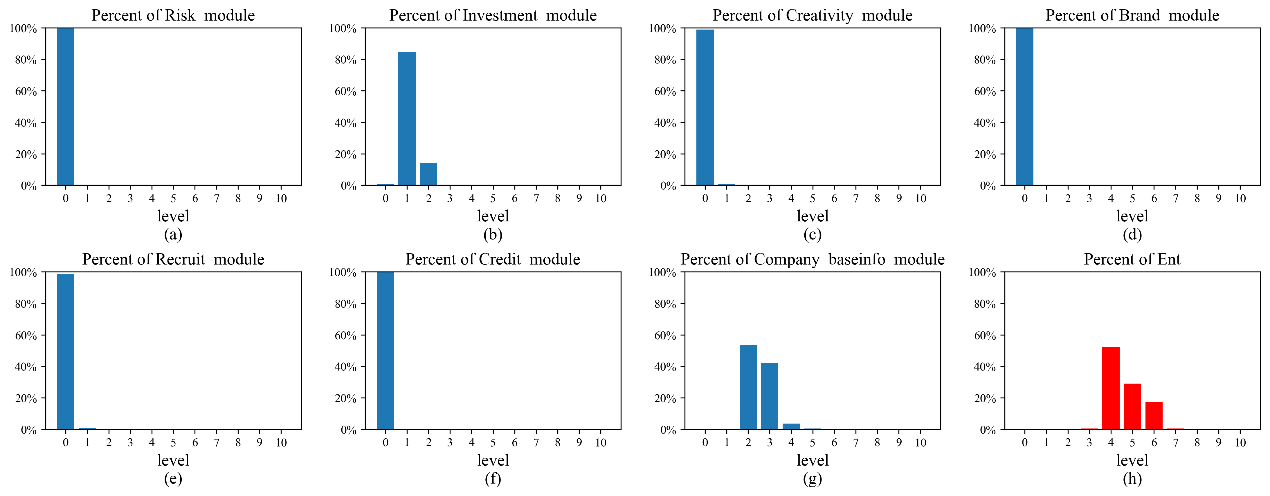


图4-4-5：企业总评4号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

如图4-4-5所示，4号簇标签：较低的投资水平，中等偏低的资产等级；企业总体水平中等。

* 5号簇：共8835条记录。

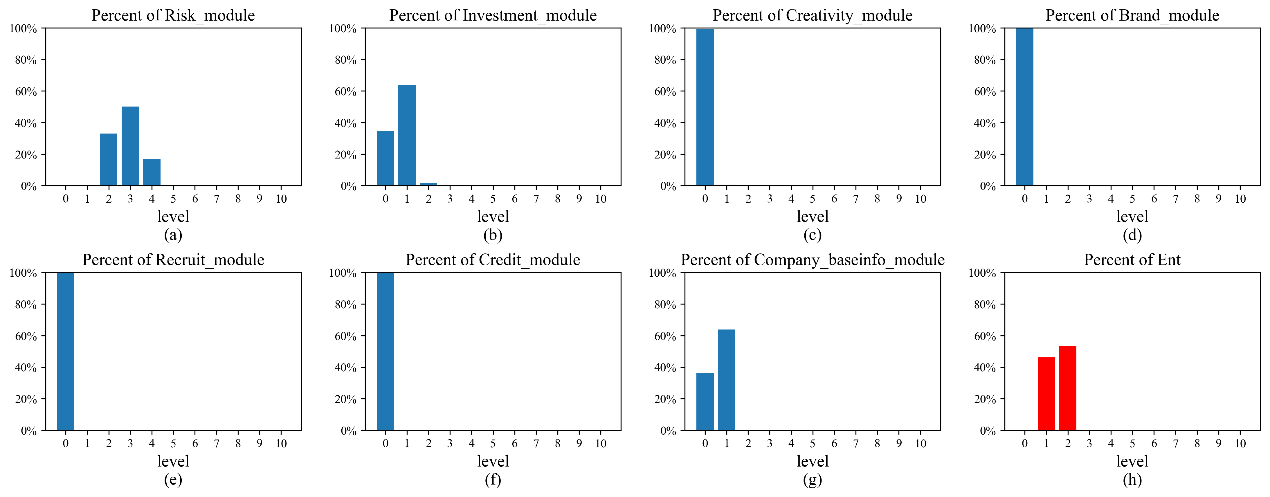


图4-4-6：企业总评5号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-6可知，5号簇标签：存在中等偏低的风险，较低的投资水平，较低的资产等级；总体企业水平较低。

* 6号簇：共6906条记录。

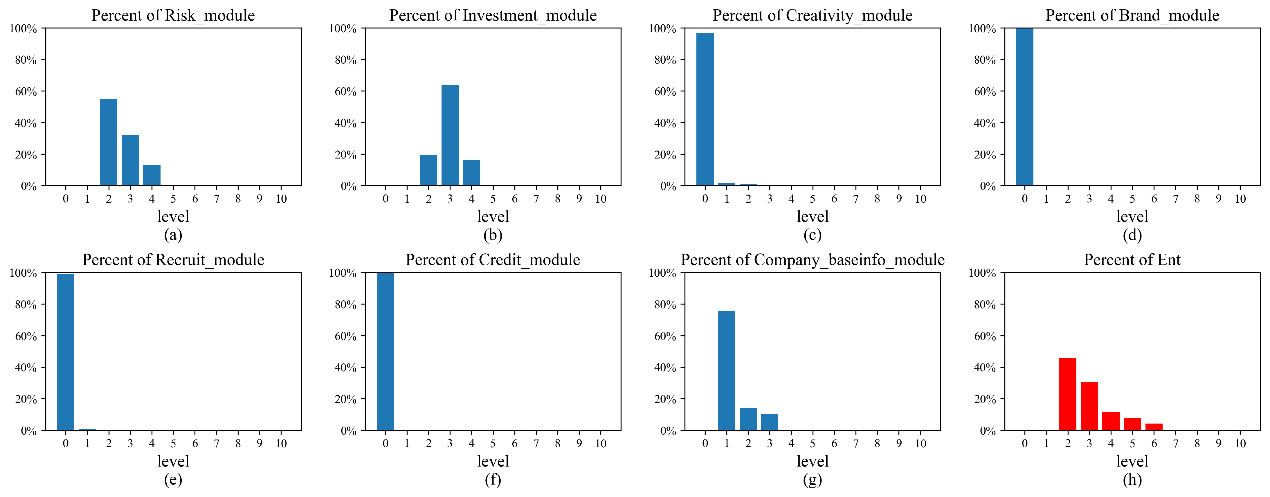


图4-4-7：企业总评6号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-7可知，6号簇标签：中等偏低风险等级，中等偏低投资水平，中等偏低企业资产等级；总体企业水平中等偏低。

* 7号簇：共26336条记录。

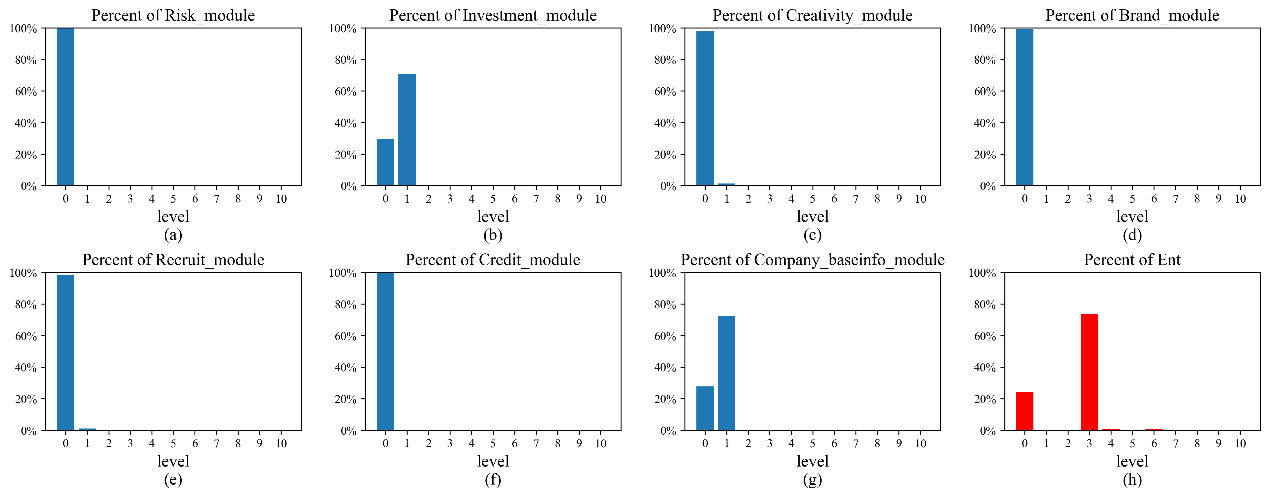


图4-4-8：企业总评7号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-8可知，7号簇标签：低投资等级，低企业资产等级；总体企业水平中等偏低。

* 8号簇：共4225条记录。

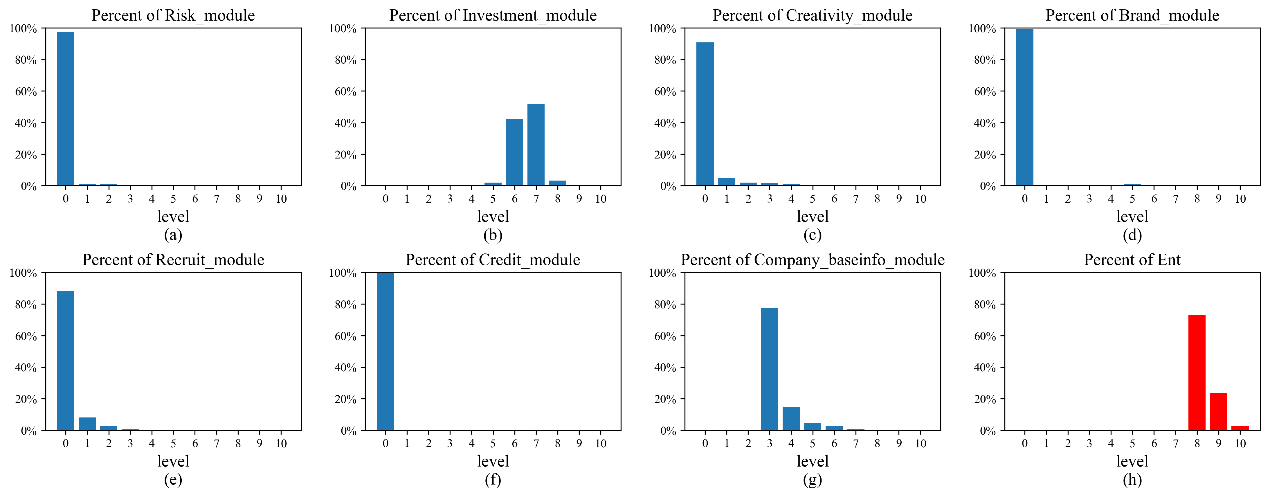


图4-4-9：企业总评8号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-9可知，8号簇标签：投资等级中等偏高，企业资产中等偏低；总体企业水平中等偏高。

* 9号簇：共5116条记录。

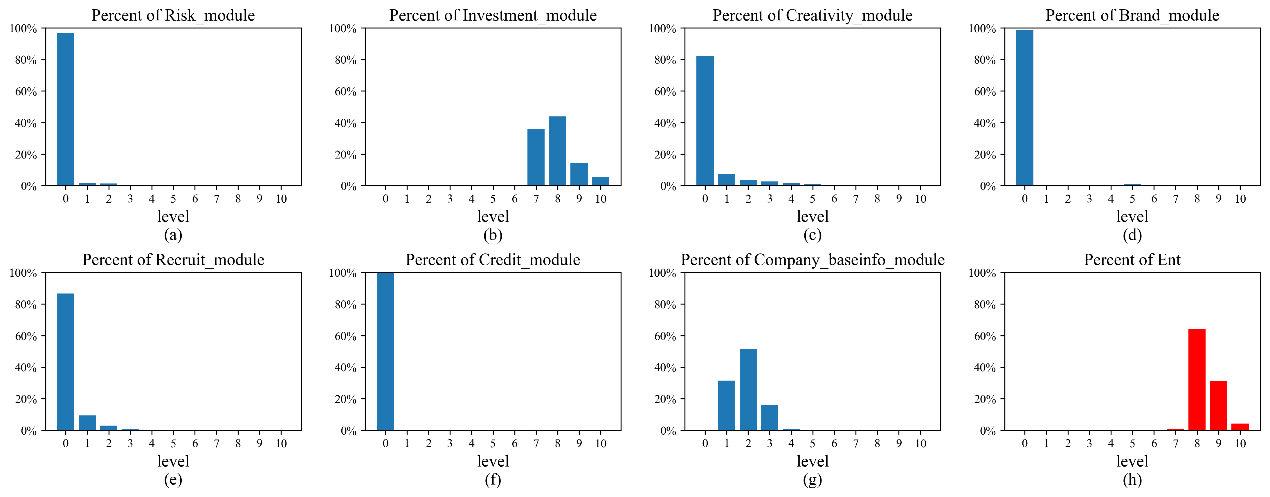


图4-4-10：企业总评9号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-10可知，9号簇标签：投资等级中等偏高，知识产权水平中等偏低，招聘等级较低，资产等级中等偏低；总体企业水平高。

* 10号簇：共2762条记录。

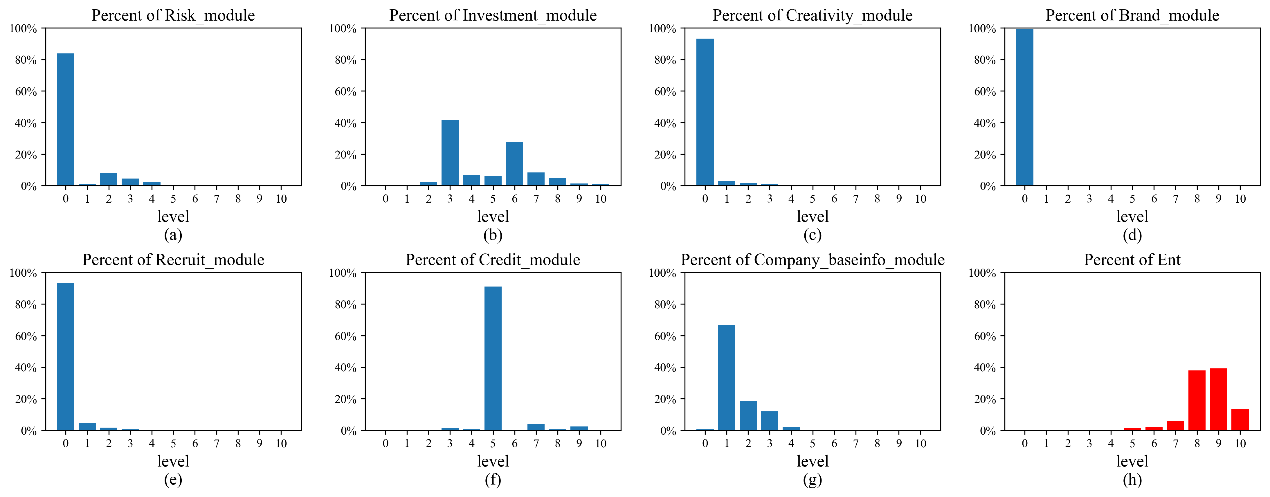


图4-4-11：企业总评10号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-11可知，10号簇标签：低风险等级，中等投资等级，中等信用，中等偏低的资产等级；总体企业水平偏高。

* 11号簇：共17468条记录。

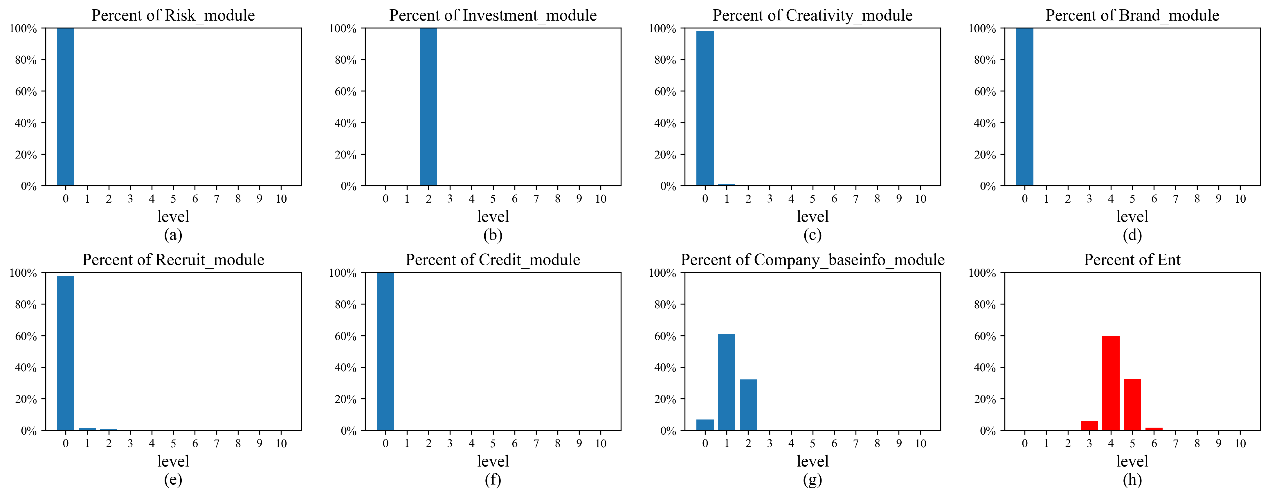


图4-4-12：企业总评11号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-12可知，11号簇标签：低投资等级，低资产等级；总体企业水平中等。

* 12号簇：共12917条记录。

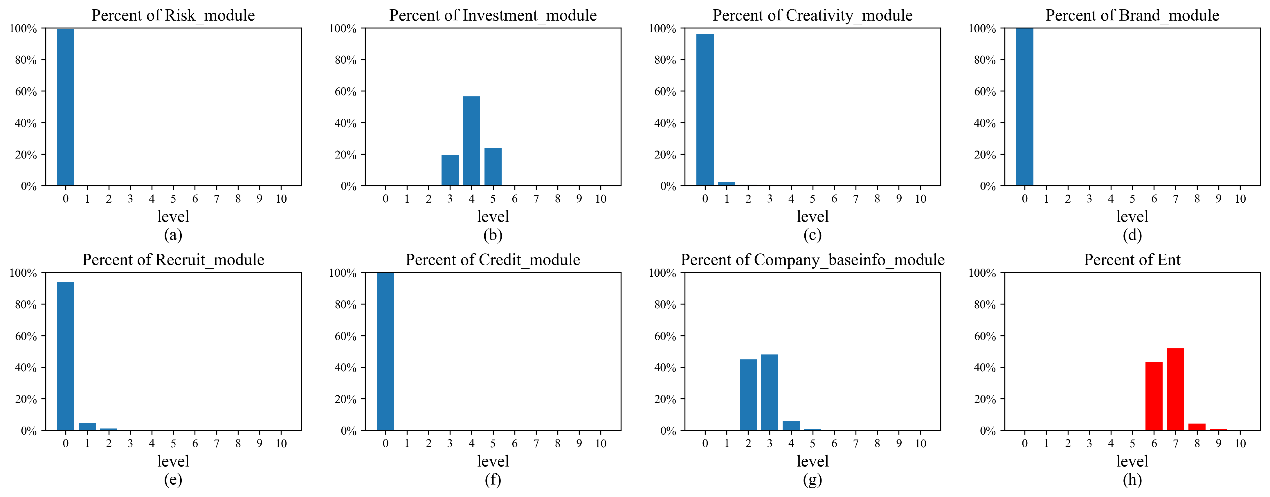


图4-4-13：企业总评12号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-13可知，12号簇标签：低投资等级，低资产等级；中等投资水平，中等偏低资产等级；总体企业水平中等偏高。

* 13号簇：共3025条记录

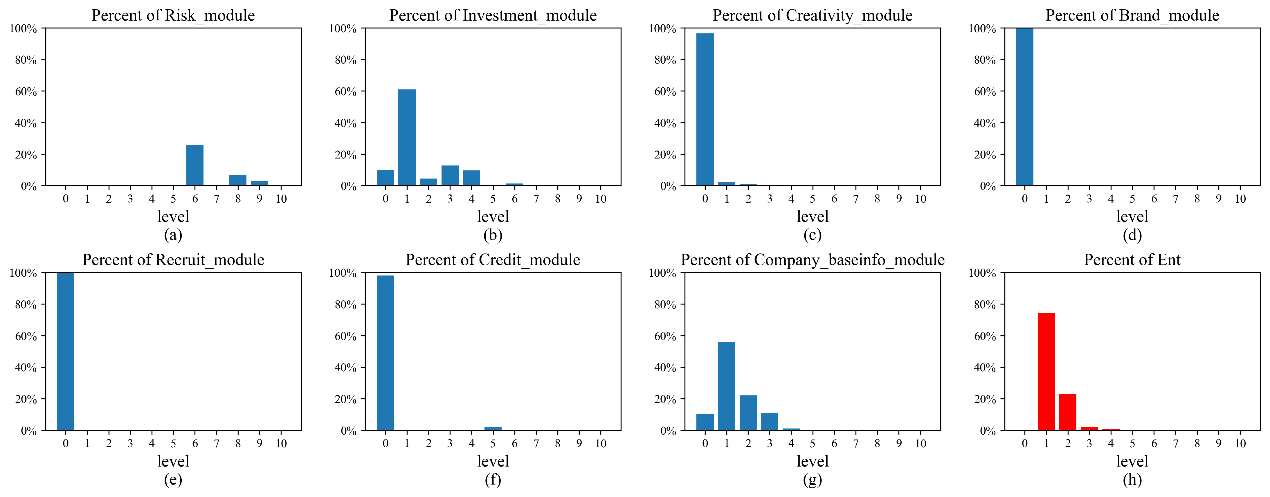


图4-4-14：企业总评13号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-14可知，13号簇标签：高风险等级，中等偏低投资等级，中等偏低资产等级；总体企业水平中等偏低。

* 14号簇：共3382条记录

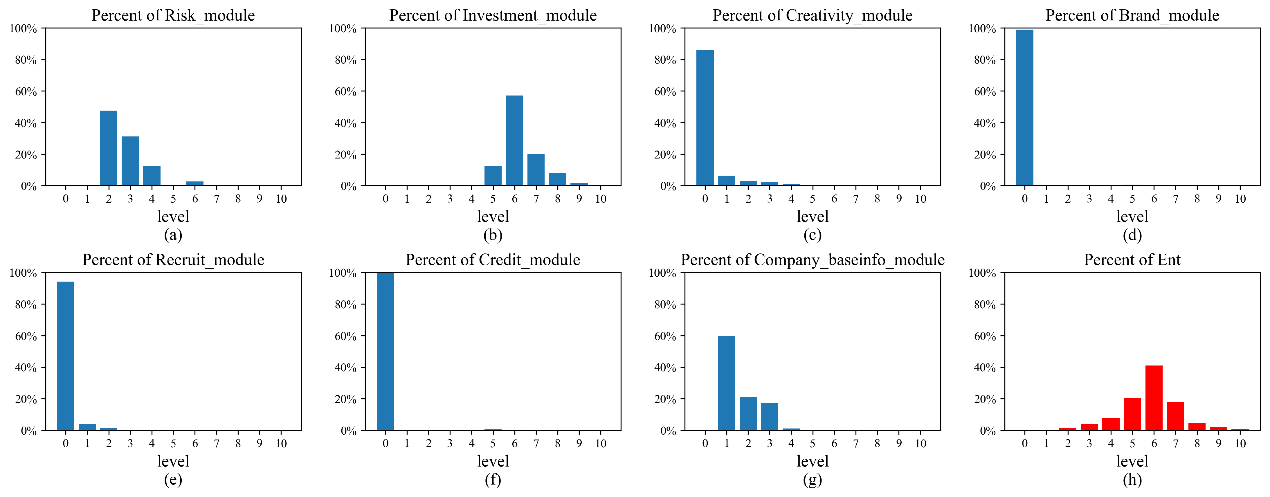


图4-4-15：企业总评14号簇各属性分布(a)风险等级分布；(b)投资等级分布；(c)知识产权等级分布；(d)品牌等级分布；(e)招聘等级分布；(f)信用等级分布；(g)基本信息(规模)等级分布；(h)企业总体等级分布

由图4-4-15可知，14号簇标签：中等风险等级，较高投资等级，具有知识产权等级、招聘等级，较低的资产等级；总体企业水平中等。

#### 4.2.4算法效率

算法的效率，主要包括预处理数据的效率，聚类模型训练效率，以及模型预测效率。

在测试算法效率过程中，使用的计算机配置如下:

操作系统：Windows10专业版64位

CPU：Intel(R) Core(TM) i7-8750H CPU @2.20GHz(12 CPUs),~2.2GHz

内存：16GB

##### 4.2.4.1预处理效率

预处理数据的过程中，由于使用K-Means算法，对企业中的属性进行了等级化操作，以统一量纲，所以预处理，也分属性的等级模型训练和模型预测时间。

对于七个模块中的表的属性，以及最终企业总评，共8张表。其中涉及到的等级化预处理的表，效率统计结果见表4-2-1到4-2-8。其中所有涉及到的时间，单位均为秒(s)。

* 风险模块

表4-2-1：风险模块预处理效率统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级化属性 | 记录数 | 训练总时间 | 预测总时间 | 训练平均时间 | 预测平均时间 |
| is\_punish | 483 | 7.95×10-3 | 2.54×10-3 | 1.65×10-5 | 5.25×10-6 |
| is\_bra | 4498 | 1.50×10-2 | 9.97×10-4 | 3.33×10-6 | 2.22×10-7 |
| pledgenum | 38 | 9.83×10-3 | 9.65×10-4 | 2.58×10-4 | 2.54×10-6 |
| is\_brap | 907 | 1.35×10-3 | 1.28×10-4 | 1.47×10-5 | 1.41×10-7 |
| taxunpaidnum | 375 | 1.40×10-2 | 9.62×10-4 | 3.72×10-5 | 2.57×10-6 |
| unpaid\_sum | 4 | / | / | / | / |
| is\_except | 28016 | 9.80×10-2 | 2.88×10-3 | 3.50×10-6 | 1.03×10-7 |
| declaredate | 119 | 1.39×10-2 | 4.52×10-4 | 1.17×10-4 | 3.80×10-6 |
| appellant\_amount | 119 | 8.65×10-3 | 8.54×10-4 | 7.27×10-5 | 7.17×10-6 |
| defendant\_amount | 119 | 1.20×10-2 | 1.00×10-3 | 1.01×10-4 | 8.40×10-6 |
| enforce\_amount | 64 | 1.09×10-2 | 6.64×10-4 | 1.71×10-4 | 1.04×10-5 |
| record\_date | 64 | 1.30×10-2 | 9.69×10-4 | 1.03×10-3 | 1.51×10-5 |
| judge\_new\_count | 998 | 1.30×10-2 | 6.57×10-4 | 1.30×10-5 | 6.58×10-7 |
| is\_justice\_credit | 976 | 1.49×10-2 | 2.01×10-4 | 1.53×10-5 | 2.06×10-7 |
| is\_justice\_creditaic | 42 | 4.11×10-3 | 2.16×10-4 | 9.79×10-5 | 5.13×10-6 |
| risk\_module | 33505 | 1.54×10-1 | 2.62×10-3 | 4.58×10-6 | 7.82×10-8 |
| 风险小计 | 33505 | 3.91×10-1 | 1.61×10-2 | 1.17×10-5 | 4.81×10-7 |

其中upaidsum由于记录数太少，所以不列入计算时间。

* 投资模块

表4-2-2：投资模块预处理效率统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级化属性 | 记录数 | 训练总时间 | 预测总时间 | 训练平均时间 | 预测平均时间 |
| insurance\_num | 42808 | 1.44×10-1 | 4.27×10-3 | 3.40×10-6 | 9.98×10-8 |
| bidnum | 1491 | 2.10×10-2 | 4.65×10-4 | 1.41×10-5 | 3.12×10-7 |
| branchnum | 156 | 1.51×10-2 | 5.42×10-4 | 9.71×10-5 | 3.48×10-6 |
| subconam\_total | 149880 | 8.37×10-1 | 1.45×10-2 | 5.58×10-6 | 9.66×10-8 |
| liacconam | 105627 | 4.83×10-1 | 9.70×10-1 | 4.57×10-6 | 9.19×10-8 |
| lisubconam | 105627 | 5.18×10-1 | 9.62×10-3 | 4.91×10-6 | 9.10×10-8 |
| investnum | 486 | 1.28×10-2 | 5.47×10-4 | 2.64×10-5 | 1.12×10-6 |
| shopnum | 3806 | 1.90×10-2 | 3.13×10-4 | 5.00×10-6 | 8.23×10-8 |
| investment\_module | 151855 | 8.55×10-1 | 1.58×10-2 | 5.63×10-6 | 1.04×10-7 |
| 投资小计 | 151855 | 2.90×100 | 4.58×10-2 | 1.91×10-5 | 3.02×10-7 |

* 知识产权模块

表4-2-3：知识产权模块预处理效率统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级化属性 | 记录数 | 训练总时间 | 预测总时间 | 训练平均时间 | 预测平均时间 |
| ibrand\_num | 446 | 9.93×10-3 | 9.98×10-4 | 2.23×10-5 | 2.24×10-6 |
| icopy\_num | 228 | 1.06×10-2 | 1.01×10-3 | 4.63×10-5 | 4.44×10-6 |
| ipat\_num | 713 | 1.40×10-2 | 8.56×10-4 | 1.96×10-5 | 1.20×10-6 |
| idom | 4858 | 2.19×10-2 | 1.66×10-3 | 4.51×10-6 | 3.43×10-7 |
| creativity\_module | 5690 | 2.39×10-2 | 3.83×10-4 | 4.27×10-6 | 6.83×10-8 |
| 知识产权小计 | 5690 | 8.03×10-2 | 3.41×10-3 | 1.41×10-5 | 5.99×10-7 |

* 品牌模块

表4-2-4：品牌模块预处理效率统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级化属性 | 记录数 | 训练总时间 | 预测总时间 | 训练平均时间 | 预测平均时间 |
| is\_jnsn | 12 | 5.17×10-3 | 3.57×10-4 | 4.31×10-4 | 2.98×10-5 |
| level\_rank | 11 | 4.99×10-3 | 1.98×10-4 | 4.54×10-4 | 1.80×10-5 |
| is\_infoa | 2 | / | / | / | / |
| is\_infob | 68 | 5.28×10-3 | 6.47×10-4 | 7.76×10-5 | 9.51×10-6 |
| passpercent | 447 | 1.20×10-2 | 3.45×10-4 | 2.67×10-5 | 7.72×10-7 |
| brand\_module | 524 | 1.59×10-2 | 5.56×10-4 | 3.03×10-5 | 1.06×10-6 |
| 品牌小计 | 524 | 4.33×10-2 | 2.10×10-3 | 8.26×10-5 | 4.01×10-6 |

由于is\_infoa记录数太少，所以不计入计算时间。

* 招聘模块

表4-2-5：招聘模块预处理效率统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级化属性 | 记录数 | 训练总时间 | 预测总时间 | 训练平均时间 | 预测平均时间 |
| qcwynum | 1371 | / | / | / | / |
| zhycnum | 209 | / | / | / | / |
| zlzpnum | 2344 | / | / | / | / |
| recruit\_module | 4970 | 4.29×10-2 | 5.18×10-4 | 8.63×10-6 | 1.04×10-7 |
| 招聘小计 | 4970 | 4.29×10-2 | 5.18×10-4 | 8.63×10-6 | 1.04×10-7 |

由于招聘模块，是直接将三家平台的招聘数求和再训练，故没有训练和预测时间。

* 信用模块

表4-2-6：信用模块预处理效率统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级化属性 | 记录数 | 训练总时间 | 预测总时间 | 训练平均时间 | 预测平均时间 |
| is\_kcont | 167 | 6.67×10-3 | 1.53×10-4 | 4.00×10-5 | 9.17×10-7 |
| credit\_grade | 29039 | 8.01×10-2 | 2.41×10-3 | 2.76×10-6 | 8.29×10-8 |
| credit\_module | 29203 | 9.60×10-2 | 2.57×10-3 | 3.29×10-6 | 8.82×10-8 |
| 信用小计 | 29203 | 1.83×10-1 | 5.13×10-3 | 6.27×10-6 | 1.76×10-7 |

* 基本信息模块

表4-2-7：基本信息模块预处理效率统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级化属性 | 记录数 | 训练总时间 | 预测总时间 | 训练平均时间 | 预测平均时间 |
| regcap | 160736 | 1.04×100 | 1.61×10-2 | 6.45×10-6 | 1.00×10-7 |
| empnum | 7.51×10-1 | 1.71×10-2 | 4.67×10-6 | 1.07×10-7 |
| esdate | 1.14×100 | 1.67×10-2 | 7.11×10-6 | 1.04×10-7 |
| baseinfo\_module | 8.67×10-1 | 1.63×10-2 | 5.39×10-6 | 1.10×10-7 |
| 基本信息小计 | 160736 | 3.80×100 | 6.61×10-2 | 2.36×10-5 | 4.11×10-7 |

* 企业总评模块

表4-2-8：企业总评模块预处理效率统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级化属性 | 记录数 | 训练总时间 | 预测总时间 | 训练平均时间 | 预测平均时间 |
| ent\_module | 187282 | 1.99×100 | 3.06×10-2 | 1.06×10-5 | 1.63×10-7 |
| 企业总评小计 | 187282 | 1.99×100 | 3.06×10-2 | 1.06×10-5 | 1.63×10-7 |

* 时间汇总

训练总时间为各个模块训练总时间求和，预测总时间即各个模块预测总时间求和；训练平均时间，即各个模块训练平均时间求和，预测平均时间，即各个模块预测平均时间求和。

记录数：189038

训练总时间：9.43×100(s)

预测总时间：1.70×10-1(s)

训练平均时间：1.77×10-4(s)

预测平均时间：6.25×10-6(s)

##### 4.2.4.2企业聚类效率

企业聚类，即用K-Means算法，对7个维度的企业数据使用进行聚类，统计其训练和预测时间。统计结果如下：

记录数：189038条；

训练总时间：2.01×100 (s)

预测总时间：3.25×10-2(s)

训练平均时间：1.07×10-5(s)

预测平均时间：1.73×10-7(s)

##### 4.2.4.3算法效率小结

最终整个模型指标以及标签建立的总时间，即数据预处理时间加企业聚类时间之和。

记录数：189038条；

训练总时间：1.14×101 (s)

预测总时间：2.03×10-1(s)

训练平均时间：6.09×10-5(s)

预测平均时间：1.08×10-6(s)

此外，在系统中测试了整个数据预处理，即训练指标提炼、到模型训练或预测所需的总时间如下：

系统训练响应时间：25.1(s)

系统预测响应时间：9.1(s)

## 五、结论

针对企业数据量大、维度丰富、缺失程度大的问题，本文通过对企业数据进分预处理后，分成7个模块，并进行加权求和的方式，有效解决了数据降维和企业数据缺失大的问题。再通过等级化的方式，统一量纲，直观反映了企业该项数据在市场上的水平。

通过等级化标签，选择了三种原理不同的模型进行对比，确定K-Means作为聚类模型，再通过不同簇的个数，综合CH、DB、SH指数进行判断，确定了最佳簇个数，使得企业的每个簇之间形成较为明显的划分界限，并给出了对应描述。

在算法效率方面，依赖python中numpy的向量化计算，以及模型的简洁明了，在训练和预测方面都有较为优秀的表现，稳定性也较强。

然而该企业分类算法也存在一定不足，例如面对企业大数据，数据可能数以亿计，然而现存的无监督训练模型，例如sklearn的K-Means，不支持分批次训练，这意味着当数据的大小超过内存加虚存时，只能通过随机抽样的方式进行训练，可能会损失一定的准确性。此外，本企业不同的属性提供的数据量大小层次不齐，有些数据量过小，不能确保等级化模型较高的准确性和稳定性。

此外，针对K-Means的使用上，仍有优化途径。如果采用NoSQL的方式存储数据结构，并使用Hadoop大数据平台进行K-Means的并行运算，效率会有进一步的提升。此外也可以尝试采用K-Means++，IsoData等K-Means改良算法，对现有聚类算法进行一定改进。

## 六、参考文献

[1] 杨俊闯,赵超.K-Means聚类算法研究综述[J].计算机工程与应用,2019,55(23):7-14+63.

[2] 田娟,朱定局,杨文翰.基于大数据平台的企业画像研究综述[J].计算机科学,2018,45(S2):58-62.

[3] 朱连江,马炳先,赵学泉.基于轮廓系数的聚类有效性分析[J].计算机应用,2010,30(S2):139-141+198.

[4] 张莉,孙钢,郭军.基于K-均值聚类的无监督的特征选择方法[J].计算机应用研究,2005(03):23-24+42.

[5] 仲兆满,管燕,胡云,李存华.基于背景和内容的微博用户兴趣挖掘[J].软件学报,2017,28(02):278-291

[6] 寇进科. 基于大数据的企业用户画像系统的设计[D].天津大学,2018.

[7] 聚类算法综述[J]. 伍育红. 计算机科学. 2015(S1)

[8] An introduction to variable and feature selection. Isabelle Guyon,Anure Elisseeff. Journal of Machine Learning Research . 2003

[9] BIRCH:An efficient data clustering method for very large databases. Zhang T,Ramakrishnan R,Linvy M. Proc.of the ACM SIGMOD Int‘l Conf.on Management of Data . 1996

[10] Optimized Data Fusion for Kernel k-Means Clustering. Shi Yu,Leon Tranchevent,Xinhai Liu,Wolfgang Glanzel,Johan A. K. Suykens,Bart De Moor,Yves Moreau. IEEE Transactions on Pattern Analysis and Machine Intelligence . 2012