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

预处理方法

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、 主题分类

由于命题方提供的数据种类丰富，为了对其更好的归类和集中处理，所以首先根据数据所描述的类型，分成了下述的 7 个模块：风险、投资、知识产权（创新能力）、品牌、招聘、信用、基本信息，并描述了其数据库的建立标准。

其中对于每张表，有以下说明事项：

- 所有的表都有属性 id，类型为 int，自增主键作为该表的唯一标识，并使用聚簇索引。
- 所有的表都有属性 entname（企业），类型为 varchar(255)，not null，在下面基本表的描述中不再赘述。
- 日期格式全部统一为 YYYY/MM/DD 格式，如果是 UNIX 时间戳的需要提前转化后导入。

1.1 风险（risk_module）

表 1-1 所示是企业风险模块中每张表以及其对应属性的解释。

表 1-1：risk_module 基本表

表名	属性名	类型	特殊要求	备注
administrative_punishment	is_punish	int	默认值 0	公司行政处罚次数
business_risk_abnormal	is_bra	int	默认值 0	列入经营异常次数
business_risk_all_punish	is_brap	int	默认值 0	行政处罚记录次数
business_risk_taxunpaid	taxunpaidnum	double	默认值 0.0	企业累计欠税额
business_risk_rightpledge	pledgenum	int	默认值 0	企业股权出质次数
ent_social_security	unpaidsocialins_so110	double	默认值 0.0	单位参加城镇职工基本养老保险累计欠缴金额
	unpaidsocialins_so210	double	默认值 0.0	单位参加失业保险累计欠缴金额
	unpaidsocialins_so310	double	默认值 0.0	单位参加职工基本医疗保险累计欠缴金额
	unpaidsocialins_so410	double	默认值 0.0	单位参加工伤保险累计欠缴金额
	unpaidsocialins_so510	double	默认值 0.0	单位参加生育保险累计欠缴金额
	updatetime	date	可为空	更新日期

exception_list	is_except	int	默认值 0	异常次数
justice_declare	declaredate	date	YYYY/MM/DD	公告时间
	appellant	int	默认值 0	上诉方(企业如果为上诉方，值为 1，否则值为 0）
	defendant	int	默认值 0	被诉方（企业如果为被诉方，值为 1，否则值为 0）
	declarestyle	varchar(255)	可为空	公告类型
justice_enforced	record_date	date	YYYY/MM/DD	执行日期
	enforce_amount	double	默认值 0.0	执行标的金额
justice_judge_new	time	varchar(255)	可为空	时间
	title	varchar(255)	可为空	标题
	casetype	varchar(255)	可为空	案件类型
	judgeresult	text	可为空	判决结果，字符较多
	casecause	varchar(255)	可为空	案由
	evidence	varchar(255)	可为空	案由编码类型
	courtrank	varchar(255)	可为空	依据
	datatype	varchar(255)	可为空	法院等级
	latypes	varchar(255)	可为空	司法类型
justice_credit	is_justice_credit	int	默认值 0	是否列入失信黑名单
justice_credit_aic	is_justice_creditaic	int	默认值 0	是否工商部失信企业

1.2 投资（investment_module）

表 1-2 所示是企业投资模块每张表以及其对应属性的解释。

表 1-2：investment_module 基本表

表名	属性名	类型	特殊要求	备注
ent_bid	bidnum	int	默认值 0	中标次数
ent_branch	branchnum	int	默认值 0	企业分支数
ent_contribution	invtype	varchar(255)	可为空	投资人类型
	conform	varchar(255)	可为空	出资方式
	subconam	double	默认值 0.0	认缴出资额
	conprop	int	默认值 0	持股比例
	condate	date	可为空	出资日期
ent_contribution_	subconcurrency	varchar(255)	可为空	认缴币种

year	accondate	date	可为空	实缴出资时间
	subconform	varchar(255)	可为空	认缴出资方式
	anchetype	varchar(255)	可为空	行业分类
	subcondate	date	可为空	认缴出资时间
	acconcurrency	varchar(255)	可为空	实缴币种
	acconform	varchar(255)	可为空	实缴出资方式
	liacconam	double	默认值 0.0	累计实缴额
	lisubconam	double	默认值 0.0	累计认缴额
ent_guarantee	priclaseckind	double	默认值 0.0	主债权种类
	pefperfrom	date	可为空	履行债务的期限自
	iftopub	varchar(255)	可为空	是否公示此担保信息 1 是 2 否
	priclasecam	double	默认值 0.0	主债权数额
	pefperto	date	可为空	履行债务的期限至
	guaranperiod	int	默认值 0	保证的期间 1 期限 2 未约定
	gatype	int	默认值 0	保证的方式 1 一般保证 2 连带保证 3 未约定
	rage	varchar(255)	默认值“0”	保证担保的范围
ent_investment	investnum	int	默认值 0	投资次数
ent_onlineshop	shopnum	int	默认值 0	网店个数
enterprise _insurance	cbrq	date	可为空	参保日期
	xzbz	varchar(255)	可为空	险种标志
	sbjgbh	varchar(255)	可为空	社会保险经办机构
	xzbzmc	varchar(255)	可为空	险种标志名称
	cbzt	varchar(255)	可为空	参保状态
	cbztmc	varchar(255)	可为空	参保状态名称
	dwbh	varchar(255)	可为空	单位编号

1.3 知识产权（creativity_module）

表 1-3 所示是知识产权模块每张表以及其对应属性的解释。

表 1-3：creativity_module 基本表

表名	属性名	类型	特殊要求	备注
intangible _brand	ibrand_num	int	默认值 0	知识产权--商标申请次数
intangible	icopy_num	int	默认值 0	企业软件著作权登

_copyright				记次数
intangible_patent	ipat_num	int	默认值 0	企业专利申请次数
web_record_info	idom_num	int	默认值 0	企业是否拥有域名的知识产权

1.4 品牌（brand_module）

表 1-4 所示是品牌模块每张表以及其对应属性的解释。

表 1-4：brand_module 基本表

表名	属性名	类型	特殊要求	备注
jn_special_new_info	is_jnsn	int	默认值 0	是否是济南市专精特新中小企业（缺失值 99.99%）
jn_tech_center	level_rank	int	默认值 0	级别(省级 2、市级 1、企业名称不出现在该表则值为 0)
trademark_infoa	is_infoa	int	默认值 0	是否列为驰名商标
trademark_infob	is_infob	int	默认值 0	是否列为著名商标
product_checkinfo_connect	passpercent	double	默认值 0.0	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）

1.5 招聘（recruit_module）

表 1-5 所示是企业招聘模块每张表以及其对应属性的解释。

表 1-5：recruit_module 基本表

表名	属性名	类型	特殊要求	备注
recruit_module	qcwynum	int	默认值 0	前程无忧招聘数
	zhycnum	int	默认值 0	中华英才招聘数
	zlzpnum	int	默认值 0	智联招聘招聘数

1.6 信用（credit_module）

表 1-6 所示是企业信用模块每张表以及其对应属性的解释。

表 1-6：credit_module 基本表

表名	属性名	类型	特殊要求	备注
enterprise _keep_contract	is_kcont	int	默认值 0	是否列为守合同重信用企业
jn_credit_info	credit_grade	varchar(255)	默认值 N	信用等级 N+、B-、A、C、N、A-

1.7 企业基本信息（company_baseinfo_module）

表 1-7 所示是企业基本信息模块每张表以及其对应属性的解释。

表 1-7：baseinfo_module 基本表

表名	属性名	类型	特殊要求	备注
company _baseinfo	regcap	double	默认值 0.0	注册资本
	empnum	int	默认值 0	从业人数
	estdate	date	默认值“0”	成立日期
	candate	date	可为空	注销时间
	revdate	date	可为空	吊销时间
	entstatus	varchar(255)	可为空	企业状态
	opto	date	可为空	经营（驻在）期限至
	enttype	varchar(255)	可为空	企业(机构)类型
	entcat	varchar(255)	可为空	企业类别
	industryphy	varchar(255)	可为空	行业门类
	regcapcur	varchar(255)	可为空	注册资本(金)币种
	industryco	varchar(255)	可为空	业务类型
	opfrom	date	可为空	经营(驻在)期限自
change_info	remark	varchar(255)	可为空	备注
	dataflag	int	可为空	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	int	可为空	变更次数
	altitem	varchar(255)	可为空	变更事项
	cxstatus	int	可为空	撤销状态： 1：变更 2：撤销变更 3：已撤销变更
	altdate	date	可为空	变更日期
	openo	varchar(255)	可为空	业务编号

二、 属性预处理

在这一步中，需要对每张表中的属性进行筛选。对于筛选出的属性，默认使用 K-Means 算法，对每一个属性（一维）进行聚类，聚成 5 类，并根据其 kmeans.center 值进行排序，相当于将各种属性统一到 5 个等级的度量。此外还有以下说明事项：

- 新生成的表默认拥有 id 和 entname 属性，要求与“主题分类”相同
- 对于“x”属性用聚类得到的标签。我们用“x_type”进行标识，并在原表附加该列。
- 对于那些属性值为空或者 0 的属性，我们采用 0 来标识。
- 对于“y”表，可能会由处理后，生成新的表，若新表无特殊含义，则用 y_p（process）表示。所有新生成的中间表用蓝色标出。
- 对于只需要用 K-Means 聚类，而不需要额外描述的表和属性，直接用“K-Means 聚类”表示其处理方式，得到 x_type。

在 2.1-2.7 中，将会根据每个表，筛选出待定的用于建模的属性，并描述它们的处理方式，和新生成的中间表。

2.1 风险预处理

表 2-1 所示是风险模块的预处理方式。

表 2-1：风险预处理

表名	有用属性	处理方式
administrative_punishment	is_punish	K-Means 聚类
business_risk_abnormal	is_bra	
business_right_pledge	pledgenum	
business_risk_all_punish	is_brap	
business_risk_taxunpaid	taxunpaidnum	一个公司可能存在多条记录，计算每个公司的欠税总和，使用 K-Means 聚类
ent_social_security	unpaidsocialins_so210-510	计算四种保险的欠款总额，得到新表 ent_social_security_p，保险欠缴总额用 unpaid_sum(int)表示，并对改属性使用 K-Means 聚类
exception_list_test	is_except	K-Means 聚类
justice_declare	declaredate，appellant，defendant	上诉方和被告方，统计每个公司上诉和被告的总次数，以及最新的纠纷日期，日期用 365*year+30*month+day 处理，生成一个新表 justice_declare_p 表，

		declaredate (int),appellant_amount(int)，defendant_amount(int)，并对三个属性进行 K-Means 聚类
justice_enforced	record_date，enforced_amount	统计出每一个公司最近一次的被执行日期（即 x=365*年+30*月+天的最大值，最终新表显示的值就是 x），以及执行金额。生成新表 justice_enforced_p，分别对 record_date 和 enforced_amount 进行 K-Means 聚类。
justice_judge_new	/	统计每个企业司法纠纷的次数，生成新表 justice_judge_new_count，和属性 judge_new_count（司法纠纷记录次数，int），使用 K-Means 聚类
justice_credit	is_justice_credit	K-Means 聚类

2.2 投资预处理

表 2-2 所示是投资模块的预处理方式。

表 2-2：投资预处理

表名	有用属性	处理方式
ent_bid	bidnum	K-Means 聚类
ent_branch	branchnum	
ent_contribution	subconam，conprop	分别是认缴出资额和持股比例。一个公司可能会有多条记录，表示多个出资人，我们希望获得一个公司的总出资额。 ①如果一条公司的多个记录当中，认缴出资额和持股比例均不为 0，则总出资额=该条记录认缴出资额/该条记录持股比例； ②如果持股比例这项的数据都是缺失的，则总出资额为所有记录认缴出资额之和 ③生成新表 ent_contribution_total，属性 subconam_total(double)，并进行 K-Means 聚类
ent_contribution_year	liacconam,lisubconam	累计实缴额和累计认缴额，同理一个公司可能出现多次，各自求出每个公司累计认缴和累计实缴的总和，然后生成新表 ent_contribution_year_total，属性名就是原来的 liacconam,

		lisubconam,并用 K-Means 聚类
ent_guarantee	/	不纳入评估
ent_investment	investnum	K-Means 聚类
ent_onlineshop	shopnum	
enterprise_insurance	cbrq	cbrq 是参保日期，根据数值规律得到前 4 位代表的是参保年份。每个企业会有多条记录，表示可能在不同年份参保了不同份数的保险。我们统计每个企业每年参保的保险平均份数。例如某公司有 6 条 2016 年的提交保险的记录，3 条 2013 年提交保险的记录，则该企业年平均提交保险份数为 4.5 份，统一向上取整汇总结果生成新表 enterprise_insurance_year_avg，属性为 insurance_num_avg (double)，并用 K-Means 聚类。

2.3 知识产权预处理

表 2-3 所示是知识产权模块的预处理方式。

表 2-3：知识产权预处理

表名	有用属性	处理方式
intangible_brand	ibrand_num	K-Means 聚类
intangible_copyright	icopy_num	
intangible_patent	ipat_num	
web_record_info	idom_num	

2.4 品牌预处理

表 2-4 所示是品牌模块的预处理方式。

表 2-4：品牌预处理

表名	有用属性	处理方式
jn_special_new_info	is_jnsn	K-Means 聚类
jn_tech_center	level_rank	
trademark_infoa	is_infoa	
trademark_infob	is_infob	
product_checkinfo_connect	passpercent	

2.5 招聘预处理

表 2-5 所示是知识产权模块的预处理方式。

表 2-5：招聘预处理

表名	有用属性	处理方式
recruit_module	qcwynum, zhycnum, zlzpnum	三家平台的招聘总数得到 recruit_sum，并做 K-Means 聚类

2.6 信用预处理

表 2-6 所示是信用模块的预处理方式。

表 2-6：信用预处理

表名	有用属性	处理方式
enterprise_keep_contract	is_kcont	K-Means 聚类

jn_credit_info	credit_grade	credit_level_dict = {
		'C-':1, 'B-':2,'A-':3,
		'A':4, 'N-':5,'N':6
		}
		将等级映射成对应的分数，使用 K-Means 聚类。

2.7 企业基本信息预处理

表 2-7 所示是基本信息模块的预处理方式。

表 2-7：企业基本信息预处理

表名	有用属性	处理方式
company_base_info	regcap，empnum，esdate	不用建立新表，但是要对这三个属性，分别 kmeans 聚类打分。此外只筛选在营企业，对它进行打分。
change_info	/	不纳入评估

三、 模块统计

在对目标属性进行聚类后，相当于是将范围不同的值，压缩到了同一个范围（1-5），这样才能更好的将多个属性，放在一起综合评估。

根据每个属性对某个模块或者主题的影响程度不同，可以设置不同的权重，进行加权求和，得到某个模块的评价总分。通过这种方式，实际上将企业的多个属性最后压缩到了 7 个维度，既尽最大可能，避免了维度压缩带来的信息损失，也避免了部分属性缺失，对评估造成的麻烦。

在模块汇总过程中，有以下说明：

- 新生成的表默认拥有 id 和 entname 属性，要求与“主题分类”相同。
- 生成模块总评即 xxx_module，即各个 module 内每个属性 type 的加权求和值。
- 每个模块总评 xxx_module 属性，都使用 K-Means 聚类，赋予 1-10 的等级，得到 xxx_module_type 属性，即模块等级。
- 对于模块内的属性，还可以通过聚类算法，找到属性间的关联，这里暂定名称为 xxx_module_inner_type（模块属性内部关联），而 inner_type 分出的类个数，将有参数调优决定。
- 合成的新表，名称为 xxx_module。

3.1 风险统计

表 3-1 所示是风险模块各个建模及等级范围、数据类型、要求以及权重。

表 3-1：risk_module

表名	属性名	类型	特殊要求	备注	权重
risk_module	is_punish_type	int	默认值 0	行政处罚等级 1-5，缺失为 0，下同	1
	is_bra_type			经营异常等级 1-5	1
	pledgenum_type			股权出质等级 1-5	1
	is_brap_type			行政处罚记录次数等级 1-5	1
	tax_unpaidnum_type			欠税等级 1-5	1.5
	unpaid_sum_type			欠缴保险额 1-5	0.7
	is_except_type			异常等级 1-5	1
	declaredate_type			最新纠纷日期 1-5	0.2
	appellant _amount_type			原告总数等级	0.3
	defendant _amount_type			被告总数等级	0.6
	enforce_amount _type			执行金额等级	1.8
	record_date _type			执行日期等级（越大等级越高）	0.3
	judge_new_count _type			诉讼次数等级	1

	is_justice _credit_type			工商部失信等级	1
	is_justice _creditaic_type			司法风险失信等级	0.6
	risk_module	double	默认值 0.0	经营风险模块加权总分	/
	risk_module_type	int	默认值 0	风险等级 1-10	/
	risk_module _inner_type			风险模块内部聚类	/

3.2 投资统计

表 3-2 所示是投资模块各个指标等级范围、数据类型、要求以及权重。

表 3-2：investment_module

表名	属性名	类型	特殊要求	备注	权重
investment _module	insurance_num_type	int	默认值 0	年平均参保等级	0.3
	bidnum_type			中标等级	2
	branchnum_type			企业分支等级	2.5
	subconam_total_type			认缴总额等级	1
	liacconam_type			累计实缴等级	0.5
	lisubconam_type			累计认缴等级	0.5
	investnum_type			投资次数等级	2.5
	shopnum _type			网店个数等级	3
	investment_module	double	默认值 0.0	投资加权总分	/
	investment_module _type	int	默认值 0	投资等级，等级 1-10	/
	investment_module _inner_type			投资模块内部聚类	/

3.3 知识产权统计

表 3-3 所示是知识产权模块各个指标等级范围、数据类型、要求以及权重。

表 3-3：creativity_module

表名	属性名	类型	特殊要求	备注	权重
creativity_module	ibrand_num_type	int	默认值 0	商标申请次数等级	1
	icopy_num_type			软著登记等级	1
	ipat_num_type			专利申请等级	1
	idom_type			域名知识产权	1

				等级	
	creativity_module	double	默认值 0.0	知识产权(创新模块)总分	/
	creativity_module_type	int	默认值 0	创新等级	/
	creativity_module_inner_type			知识产权（创新模块）内部聚类	/

3.4 品牌统计

表 3-4 所示是品牌模块各个指标等级范围、数据类型、要求以及权重。

表 3-4：brand_module

表名	属性名	类型	特殊要求	备注	权重
brand_module	is_jnsn_type	int	默认值 0	济南市中专精小企业等级	1.3
	level_rank_type			科技等级	1.5
	is_infoa_type			驰名商标	1.5
	is_infob_type			著名商标	1
	passpercent_type			质检通过率	0.7
	brand_module	double	默认值 0.0	品牌模块总分	/
	brand_module_type	int	默认值 0	品牌等级 1-10	/
	brand_module_inner_type			品牌模块内部聚类	/

3.5 招聘统计

表 3-5 所示是招聘模块各个指标等级范围、数据类型、要求以及权重。

因为三张表已经汇总过了，所以不用再额外操作，只需要直接对三大平台的招聘数做聚类打分即可，由于只对招聘总数做评估，所以不做内部属性聚类。

表 3-5：recruit_module

表名	属性名	类型	特殊要求	备注	权重
recruit_module	qcwynum	int	默认值 0	前程无忧招聘	1
	zhycnum			中华英才招聘	1
	zlzpnum			智联招聘	1
	recruit_module			三个平台招聘总数	/
	recruit_module_type			招聘数等级 1-10	/

3.6 信用统计

表 3-6 所示是信用模块各个指标等级范围、数据类型、要求以及权重。

由于守合同重信用企业的数据量相比信用等级数据量少了太多，且本身数据属性维度仅为 2 维，只对信用模块加权总分做聚类，不进行内部聚类。

表 3-6：credit_module

表名	属性名	类型	特殊要求	备注	权重
credit_module	is_kcont_type	int	默认值 0	守合同重信用企业等级	1.3
	credit_grade_type			信用等级	1
	credit_module	double	默认值 0.0	信用模块加权总分 1-10	/
	credit_module_type	int	默认值 0	信用等级	/

3.7 企业基本信息统计

表 3-7 所示是基本信息模块各个指标等级范围、数据类型、要求以及权重。

由于注册资本权重很高，内部分类实际就是企业基本信息等级分类，所以不做内部分类。

表 3-7：company_baseinfo_module

表名	属性名	类型	特殊要求	备注	权重
company_baseinfo_module	regcap_type	int	默认值 0	注册资本等级	/
	company_baseinfo_module	double	默认值 0.0	基本信息模块总评	/
	company_baseinfo_module_type	int	默认值 0	企业基本信息等级	/

四、 模块总评

经过了步骤三的模块统计后，企业的数据已经成功被压缩到了 7 个维度，接下来需要将各个 module_type 通过加权求和的方式，获得企业的总评分，并再使用 K-Means 的方式获得其总评等级（此处分为 10 个等级），并再根据 7 个维度的 module_type，直接做 K-Means 训练，挖掘每个属性之间的内部关联，获得 ent_inner_type，得到最终表 ent 如表 4-1 所示：

表 4-1：ent

表名	属性名	类型	特殊要求	备注	权重
ent	risk_module_mark	int	默认值 0	1-风险	-1.11
	investment_module_mark			2-投资	1

	creatity_module_mark			3-创新能力	1
	brand_module_mark			4-品牌	1
	recruit_module_mark			5-招聘	1
	credit_module_mark			6-信用	1
	base_info_module			7-基本信息	1
	ent	double	默认值 0.0	企业总分	/
	ent_type	int	默认值 0	企业总分等级，10 个等级	/
	ent_inner_type			分成若干个类	/