基于加权多层 K-Means 的企业分类系统

“e 企查”方案概要

一、前言

金融科技场景中，企业这一信贷主体，数据量大且来源广泛，企业信息维度丰富，在分析企业还款能力、信用水平过程中，面临巨大的挑战。

- 如何对无标识的企业数据进行预处理、特征筛选、提取，形成有效的评估指标？
- 如何对提取的指标，进行高效的无监督分类，对小微企业群体形成合适划分？
- 如何对分类的评估结果，构建企业画像和信用评分体系？
- 如何根据企业画像和信用评分，挖掘更多潜在的值得投资的企业？

在充分调研和评估企业特征信息的基础上，开发基于加权多层 K-Means 算法的企业分类系统“e 企查”。该平台的定位如下：

①旨在为企业提供精准的企业画像和信用评估，为企业生成明显的标签，为金融机构提供可视化服务和辅助评估手段。

②结合无监督分类算法设计推荐算法，挖掘潜在类似企业。

③对于企业数据管理者，并提供机器学习模型测试环境，更加及时的调整模型。

二、创意描述：

①设计加权多层 K-Means 算法，进行企业准确分类和画像构建，有效解决数据维度大、量大、缺失严重、量纲不一致的问题，准确高效生成企业标签。

②使用智能推荐算法，挖掘潜在客群。

③数据管理驾驶舱，丰富的可视化和辅助评估手段，了解和比较企业数据。

④单独的机器学习测试界面，可以调整模型参数，在应用前准确把握模型性能。

三、功能简介：

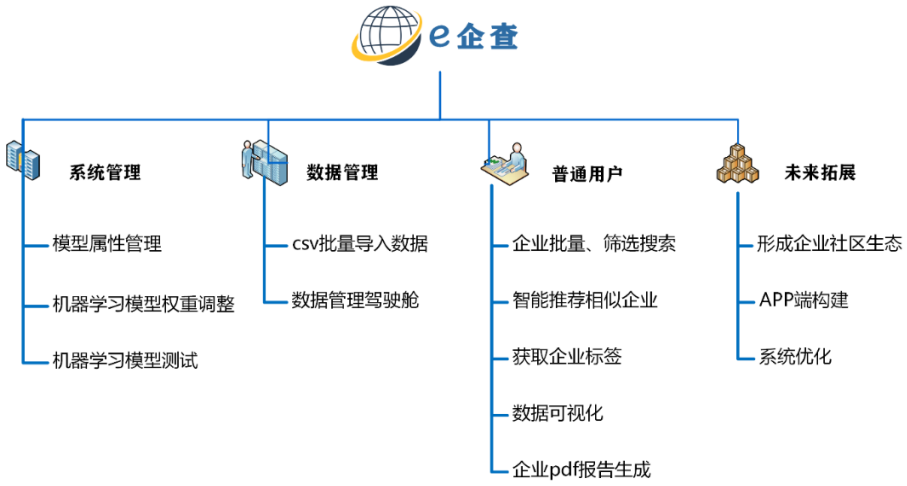

图 3-1：功能模块划分图

四、特色综述

◆ 算法部分：

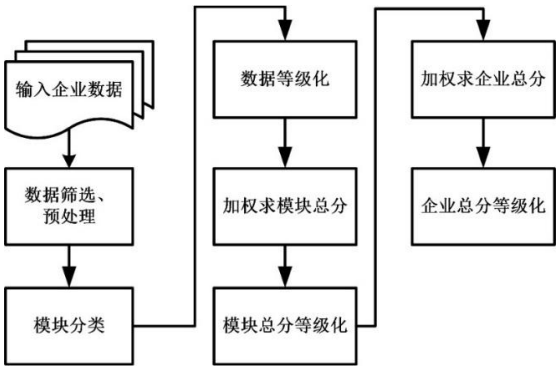

图 4-1：算法整体流程

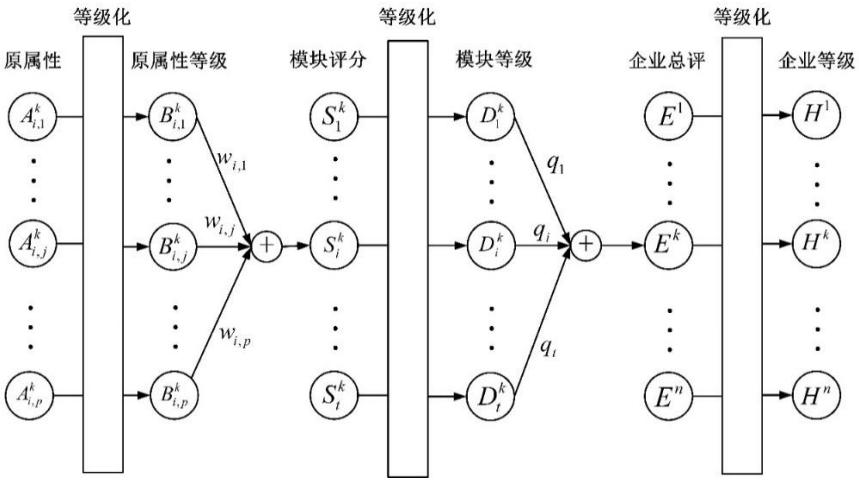

图 4-2：建模指标数据流

加权多层K-Means算法实现过程,如图4-1的算法流程和图4-2所示的企业数据流。特色如下：

①采用等级化算法，统一不同属性间量纲，高效，鲁棒性强，快速明确企业属性在整体中的水平。

②模块分类降维，将企业属性根据信息类型分成7大模块，解决数据缺失问题，提升聚类效率，能使企业在较少数据的情况下也能生成画像。

③加权K-Means聚类，调整不同属性对于最终聚类结果的影响程度。

④智能推荐算法，综合考虑行业信息、加权K-Means聚类结果，企业总评等，准确为企业推荐潜在客群。

⑤算法性能良好。在本系统中，对于出题方所提供的所有企业数据（19万条左右的企业记录），在笔记本电脑上（CPU为Intel i7-8750H）本地测试，指标如表1所示：

表1：算法处理效率

事件	总时间（s）
模型训练总时间	11.4
预测样本总时间	0.2
系统训练响应时间（预处理+训练）	25.1
系统预测响应时间（预处理+预测）	9.1

评估指标使用CH（Calinski-Harabasz）、DB（Davies-Bouldin）、SH（轮廓系数）。得到结果如下：CH= 136499.419；DB= 0.868；SH= 0.564。该方法在对比传统K-Means、Birch（层次聚类）、GaussianMixture（高斯混合聚类），无论从时间效率还是评估指标，都有较大提升。

◆ 系统功能部分：

①丰富的可视化手段，直观显示企业的各模块水平，也可以在企业之间形成对比。对于后台管理模块，模型分类打标签结果、企业整体数据分布可视化。

②模型管理。管理员可以更新者模块属性权重等，也拥有模型的单独测试界面。

③企业评估文档自动生成，辅助金融信贷机构进行投资选择。

④多种筛选审查方式，帮助企业找到心仪企业。

⑤涉及到数据操作的部分均有批量操作接口，提升使用体验。

⑤前后台分离，分工明确，给企业用户良好的交互体验的同时，也给数据和模型维护者更好的管理体验。

五、开发工具与技术

表 2：开发工具与技术

名称	版本	属性
MySQL	5.1	数据库
Django	2.2.3	后端框架
Sklearn	0.0	机器学习框架
Python	3.7.0	模型训练、后端
Vue	2.0	前端框架
PyCharm	2017	集成开发工具

六、应用对象

全国各大企业和金融信贷机构。

七、应用环境

表 3：应用环境说明

环境类型	配置说明
硬件环境	华为云主机，单核 CPU，2G 内存
数据库	MySQL 5.1
浏览器	IE、Chrome、Firefox、Edge 等主流浏览器

八、结语

e 企查，在深入了解行业同类竞品的调研以及从业人员，中小企业痛点问题，制定相应的业务需求。与传统的商业查询平台相比，我们能够更加清晰地抽象出企业特征，实现更好、更高效的企业分类。