

关于拆分项目规避招标数据模型的研究

卢亚茹

中国石油化工集团有限公司，北京 100728

摘要： 提出一种基于人类行为动力学理论、核密度聚类、文本相似度分析技术搭建的拆分项目规避招标模型，通过对合同管理系统中工程类合同的合同金额、签订时间、合同名称等信息的分析，发现企业故意拆分工程项目签订多份合同情况，为核实规避应招必招问题提供线索。

关键词： 人类行为动力学；核密度聚类；文本相似度分析；拆分项目规避招标

Research On Splitting The Project Avoidance Of Bidding Data Model

Lu Yaru

China Petroleum & Chemical Corporation Limited, Beijing 100728

Abstract： Puts forward a based on human behavior dynamics theory, nuclear density clustering, text similarity analysis technology to build the split project to avoid bidding model, through the contract management system of engineering contract amount, signing time, the name of the information such as analysis, found that the enterprise deliberately split project signed more contracts, to verify evade will provide clues.

Keywords： human behavior dynamics; kernel density clustering; text similarity analysis; split projects to avoid bidding

研究背景

在审计过程中，审计人员发现企业在工程项目中存在通过拆分项目相关费用需求，以达到规避招标的目的^[1]。这种行为存在资金风险，有必要作为审计关注点。目前，审计人员主要靠抽样的方式，人工识别相似的合同名称来找出此类问题。但由于工程项目涵盖范围广泛，各类工程费用需求种类繁多，抽样方式可能存在漏检，若要实现全量检测，靠人工的方式将会是一项非常繁重的工作^{[2][3]}。基于人群行为动力学理论、相似性聚类、NLP 文本识别技术搭建的拆分项目规避招标模型，通过对合同管理系统中工程类合同的合同金额、签订时间、合同名称等信息的分析，发现企业故意拆分工程项目签订多份合同情况，为核实规避应招必招问题提供线索。该模型能实现基于大数据样本的检测监控，减轻审计人员的工作量。

一、算法简介

本文基于人类行为动力学-幂律分布生成维度（指数尾系数，阵发/记忆系数），然后利用核密度聚类，对供应商进行聚类，利用文本相似度分析对疑似供应商合同数据进行分析。

（一）幂律分布

泊松分布：以往的研究中，泊松过程作为经典的刻画人类活动模式的手段。按照泊松原理，两个随机事件到达的间隔时间的概率密度函数 τ 具有指数形式：

$$p(\tau)=\lambda e^{-\lambda\tau}$$

幂律分布：概率密度函数服从幂函数的分布，是一种长尾分布，广泛存在经济与金融学领域。通常如果事件发生的振幅（如供应商签订合同时间间隔、金额等）不符合自然规律时，会导致事件的低频部分呈现异常分布，指数尾的计算值能有效发现这一异常^[4]。

幂律分布的事件间隔是一个幂函数形式：

$$p(\tau) \propto \tau^{-\alpha}$$

最近几年，越来越多的证据显示人们的很多行为的时间统计特性无法用泊松过程刻画。特别地，这些行为所对应的间隔时间分布具有明显偏离指数分布的肥胖的尾部，可以用幂函数更好拟合。

$$p(\tau)=\lambda e^{-\lambda\tau} \longrightarrow p(\tau) \propto \tau^{-\alpha}$$

1. 指数尾系数

根据幂律分布，找到指数分布和幂律分布的拐点，计算斜率，该斜率即为指数尾系数。

（1）画直方图。计算每个供应商参加招投标活动的时间间隔频次。

（2）找拐点。找到图形中指数分布和幂律分布的拐点

（3）算斜率。利用最小二乘法计算幂指数斜率，作为供应商一个维度

2. 阵发 / 记忆系数

阵发性：另一种定量刻画间隔分布的方法，这些方法并不需要先假设该分布满足某种特殊的函数形式。当人类行为的时间特性可以用泊松过程描述时，事件发生频率比较均匀；反过来，当时间间隔分布服从幂函数律的时候，有时候很多事件会在较短时间内密集发生，然后又会出现一个很长的空档期，把后面这种情况叫做阵发性^[5]。

引入：

$$B = \frac{\sigma_\tau - m_\tau}{\sigma_\tau + m_\tau}$$

来刻画阵发性，其中 σ_τ 和 m_τ 分别表示分布 $p(\tau)$ 的标准差和平均值， B 的取值显然在 $-1 \sim 1$ 之间。对于泊松分布，均值和标准差相等，因此阵发性恰为 0，可以看作 Goh 和 Barabási 设定的平衡点。对于严重的胖尾分布，标准差远大于均值， B 靠近 1；对于周期性规则信号，时间间隔分布实际上是一个 δ 函数，标准差为 0， B 值为 -1 。

由于人类行为在发生时间上具有“强阵发弱记忆”的特性，所以本文提取维度只考虑阵发性，统计每个供应商参加招标投标活动的金额，计算阵发性指标，得到阵发性 / 记忆系数。

3. 供应商维度提取

阵发性 / 记忆系数作为供应商另一个维度，完成供应商维度提取（指数尾系数，阵发 / 记忆系数）。

（二）核密度估计的聚类算法

1. 核密度估计

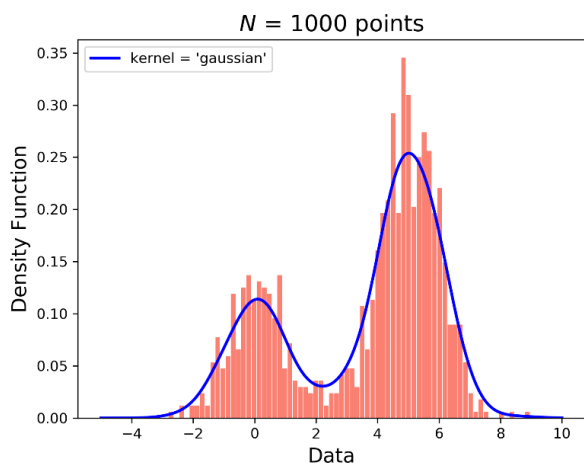
核密度估计（kernel density estimation）不利用有关数据分布的先验知识，对数据分布不附加任何假定，是一种从数据样本本身出发研究数据分布特征的方法，属于非参数检验方法之一^[6]。

核函数选择均匀核函数：

$$K(x) = \frac{1}{2}, \quad -1 \leq x \leq 1,$$

加入带宽 h 后： $K_h(x) = \frac{1}{2h}, \quad -h \leq x \leq h$

下图和频率分布直方图的比较能更好的理解核密度估计，蓝色的线为概率密度估计，橙色的柱状图为数据出现的频率统计^[6]

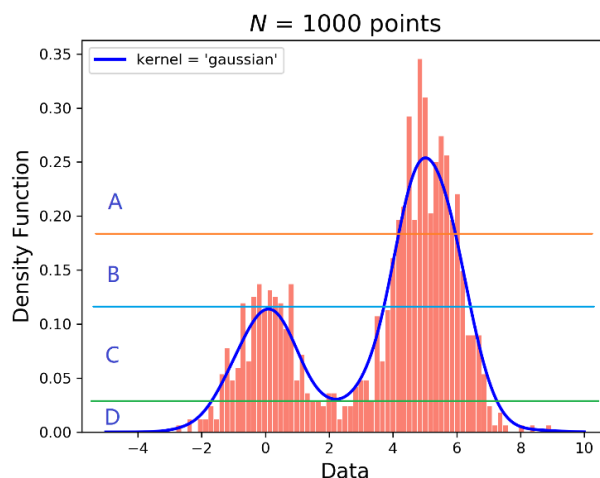


2. 核密度聚类

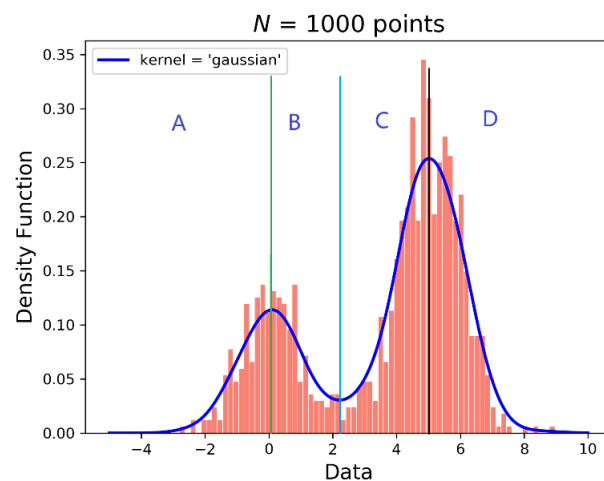
基于核密度估计的聚类算法原理是从一个任意的数据点出发，运用爬山法搜索其所属的局部吸引子。若吸引子所在点处的密度函数值达到或超过预设的阈值，则将该点赋给由吸引子生成的类，否则，认为该吸引子所吸引的数据点太少而将所吸引的样本点视为噪声数据^[7]。

两种聚类方法

①按照曲线中的极大值和极小值分成几个区间根据数据的出现频率完成 A、B、C、D 分类



②按极值点的横坐标取值将数据分为每个数据段都有不同的含义的 A、B、C、D 共四个类



（三）文本相似度分析

1. 普通文本相似度

余弦相似度就是通过一个向量空间中两个向量夹角的余弦值作为衡量两个个体之间差异的大小。把 1 设为相同，0 设为不同，那么相似度的值就是在 $[0, 1]$ 之间，所有的事物的相似度范围都应该是 $[0, 1]$ 。余弦值接近于 1，夹角趋于 0，表明两个向量越相似^[9]。

（1）分词：将每个合同名称进行分词。

（2）对比：计算每两份合同的文本相似度，

$$\cos \theta = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} = \frac{A \times B}{|A| \times |B|} \quad (1)$$

其中 A 和 B 是两份合同名称分词后的 n 维向量
鉴于合同名称的特殊性, 本文选择余弦相似度计算。

2. 专有词语相似度

- (1) 构建专有词词库;
- (2) 提取所有可识别的专有词语;
- (3) 对比: 计算每两份合同间提取的专有词相似度, 找寻最大的相似度值, 对全域求平均^[10]。

公式如下:

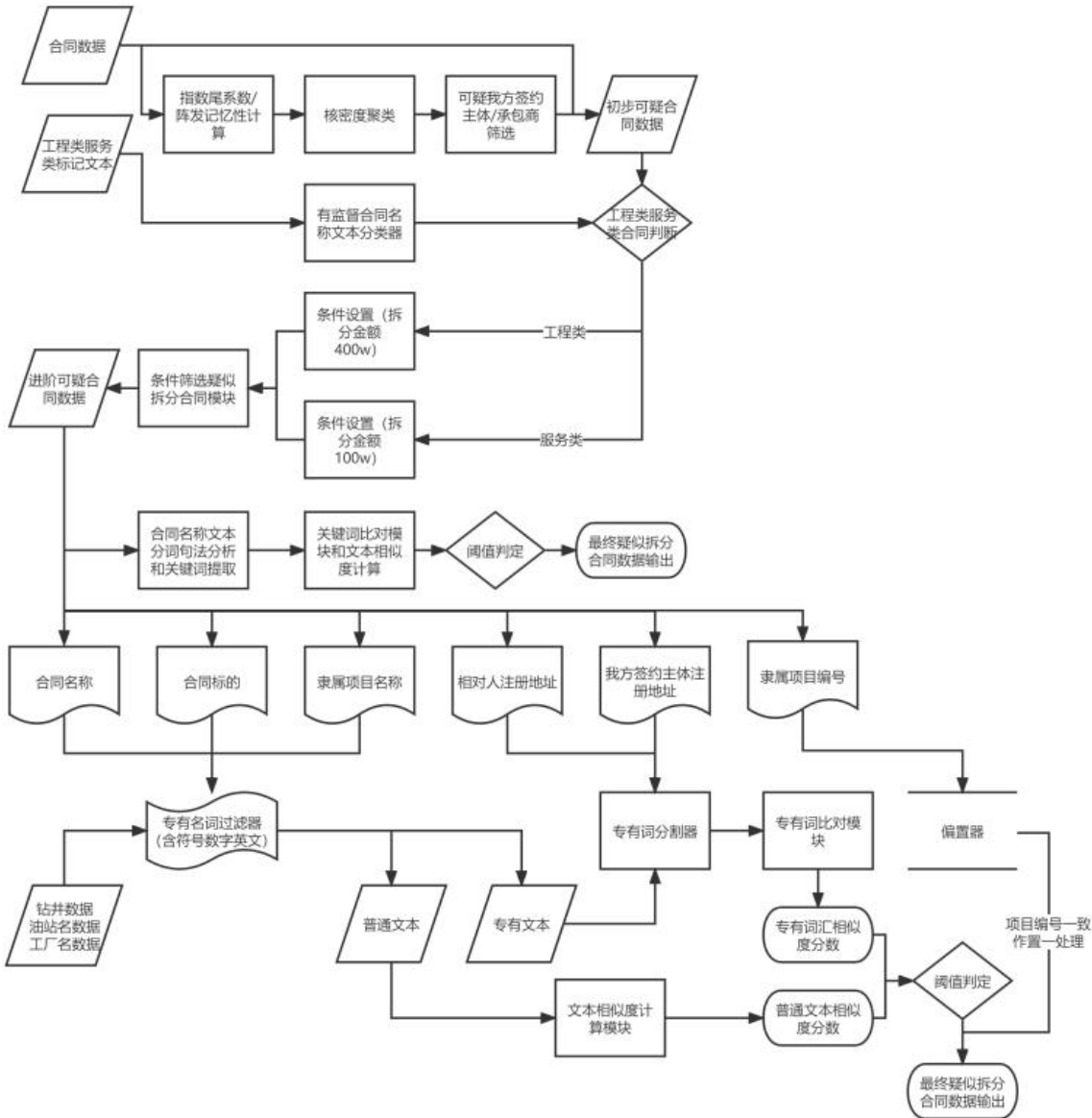
$$P(A,B)=\frac{\sum_j A_j+\sum_j B_j}{n_A+n_B} \tag{2}$$

其中 A 、 B 为合同名称的专有分词集合, n_i 为 i 集合的频次。

二、案例简介

(一) 分析流程

整个模型分析流程如下



(二) 疑似人为拆分合同的供应商识别

- (1) 聚类。将每个供应商的两组维度（指数尾系数、阵发 / 记忆系数），利用 KDE 方法进行聚类，得到聚类图
- (2) 筛查。最终计算聚类结果中各供应商位置与计算点位的偏离度，偏离度越高则该供应商出现异常的可能性越大。



公司代码	35000020
指数尾指标	-0.925553514
阵发 / 记忆性指标	0.867926965
指标偏离度	2.350677
判断	异常

公司代码	40727980
指数尾指标	-2.205477145
阵发 / 记忆性指标	0.956426119
指标偏离度	0.07846
判断	正常

正常供应商的指标表现高度相似，在聚类中应当表现在聚类的中心点，而异常的供应商则会表现出多种不同的异常情况，往往离散在中心点外围。建立供应商“合同事件”视界地图，划分“自然签约合同”和“疑似人为拆分合同”的供应商分布区域，识别疑似人为拆分合同的供应商。

（三）疑似拆分工程合同的合同组合识别

取疑似人为拆分合同的供应商的所有合同，利用文本相似度分析，筛选出疑似拆分工程合同的合同组合。

以表格中的合同为例

	合同名称	合同标的 （如有）	项目名称 （如有）	我方签约 主体工商 注册地	对方签约 主体工商 注册地
A	塑料部1/2PP装置包装设施升级改造项目建设工程设计合同	/	塑料部1/2PP装置包装设施升级改造项目	上海市金山区金一路48号	上海市金山区金一路143号

3. 输出结果并验证

公司代码 / 简称	供应商编号 / 名称	合同序号	合同编号	合同名称	合同创建日期	合同金额	合并合同 总额	普通词典 相似度	专业词典 相似度
2480	40562360	1900718808	30450000-19-FW0199-0544	塑料部4PE装置包装设施升级改造项目建设工程设计合同	2019-12-23	488,000	1,949,900	0.8825	0.75
		1900721179	30450000-19-FW0199-0546	塑料部1/2PE装置包装设施升级改造项目建设工程设计合同	2019-12-23	479,300			
		1900721468	30450000-19-FW0199-0542	塑料部1/2PP装置包装设施升级改造项目建设工程设计合同	2019-12-23	488,600			
		1900721638	30450000-19-FW0199-0545	塑料部聚烯烃装置包装设施升级改造项目建设工程设计合同	2019-12-23	494,000			

通过普通词典相似度、专业词典相似度更精准地圈定范围，根据该算法，输出结果中包含了用户提供的实例中5份合同里的4份合同，验证了该算法的可行性。

三、结论

本文提出了一种基于人类行为动力学理论、核密度聚类和文本相似度分析技术的拆分项目规避招标模型。该模型通过分析合同管理系统中的工程类合同金额、签订时间和合同名称等信息，有效地识别了企业故意拆分工程项目签订多份合同的行为，为审计人员核实规避招标问题提供了有力线索。

参考文献：

[1] 何一平；郑哈哈；詹雯；赵梦天. 论规避公开招标[J]. 中国政府采购, 2023(12):56-63.
[2] 刘涛. 从“以化整为零方式规避公开招标”说开去[J]. 中国政府采购, 2017(09):46-49.
[3] 廖秀红. 工程招中标标段划分及多标段招标的定标方式探讨[J]. 江西建材, 2022(12):431-432+437.
[4] 周涛, 韩筱璞, 闫小勇, 杨紫陌, 赵志丹, 汪秉宏. 人类行为时空特性的统计力学. 电子科技大学学报. 2013, 42 (04).
[5] 李凯凯. 用户行为记忆性及其驱动机制研究[D]. 中北大学, 2017.
[6] 朱杰, 陈黎飞. 核密度估计的聚类算法 模式识别与人工智能 PR & AR. 2017年5月 第30卷第5期.
[7] 熊开玲, 彭俊杰, 杨晓飞, 等. 基于核密度估计的 K-means 聚类优化[J]. 计算机技术与发展, 2017, 27 (02): 1-5.
[8] 韩鑫. 基于核的层次聚类算法研究[D]. 西安石油大学, 2021. DOI:10.27400/d.cnki.gxasc.2021.000833.
[9] 张振亚, 王进, 程红梅, 等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学, 2005, (09): 160-163.
[10] 王春柳, 杨永辉, 邓霏, 等. 文本相似度计算方法研究综述[J]. 情报科学, 2019, 37 (03): 158-168. DOI:10.13833/j.issn.1007-7634.2019.03.026.

	合同名称	合同标的 （如有）	项目名称 （如有）	我方签约 主体工商 注册地	对方签约 主体工商 注册地
B	塑料部聚烯烃装置包装设施升级改造项目	/	塑料部聚烯烃装置包装设施升级改造项目	上海市金山区金一路48号	上海市金山区金一路143号

1. 合同名称中的专有名词

A: { 塑料部, 1/2PP, 装置 }

B: { 塑料部, 聚烯烃, 装置 }

A: { 上海市, 金山区, 金一路, 48号 }

B: { 上海市, 金山区, 金一路, 48号 }

利用公式（2）计算相似度。

2. 合同名称中的普通文本分词

A: { 包装, 设施, 升级, 改造, 建设, 工程, 设计 }

B: { 包装, 设施, 升级, 改造 }

利用公式（1）计算相似度。

该模型的优势在于：

· 基于大数据样本：能够对大量合同数据进行全面分析，避免传统抽样方法的漏检问题。

· 自动化检测监控：减轻审计人员的工作负担，提高审计效率。

· 多维度分析：结合人类行为动力学理论、核密度聚类和文本相似度分析技术，从多个维度识别异常行为，提高识别准确率。

该模型的实际应用结果表明，其能够有效地识别疑似人为拆分合同的供应商及其合同组合，验证了模型的可行性和有效性。

未来，可以进一步研究如何将该模型应用于其他领域，例如识别虚假交易、反洗钱等，以更好地发挥其作用。