doi:10.6043/j.issn.0438-0479.202407006

基于 EM 算法与混合模型的动态聚类分析

金向阳[1],章惠民[2],王语涵[3],林建华[3*]

(1. 中国烟草总公司福建省公司,福建 福州 350003; 2. 福建省烟草公司漳州市公司,福建 漳州 363000;
3. 厦门大学数学科学学院,福建 厦门 361005)

摘要:[**目的**]对 2022 年福建漳州烟草公司品牌销售数据开展动态聚类,以揭示数据深层结构,支撑市场策略优化.[**方法**]研究综合运用 EM 算法与高斯混合模型进行参数估计及动态聚类.依托统计软件实现算法流程,包括参数初始化、EM 迭代优化及基于概率分布的聚类,严格遵循统计原则保障结果客观性.[**结果**]新算法有效估计概率模型参数,实现烟草品牌精准动态聚类.分析揭示了各品牌类别的差异化特征,为市场策略定制及产品组合优化提供依据.算法准确计算品牌在各类别中的概率分布,增强了决策的精准性.同时,算法具备灵活性与适应性,可随市场变化动态调整.[**结论**]本研究提出的基于混合高斯分布与 EM 算法的数据分析方法,为市场数据分析提供了新视角.该方法提高了数据分析的精度与效率,助力企业在复杂市场环境中制定科学策略,具有良好的应用价值与推广前景.

关键词:概率模型;EM 算法;混合分布;动态聚类

中图分类号:O 212 **文献标志码**:A **文章编号**:0438-0479(2025)04-0727-13

Dynamic clustering analysis based on the EM algorithm and mixture models

JIN Xiangyang[1], ZHANG Huimin[2], WANG Yuhan[3], LIN Jianhua[3*]

(1. Fujian Branch of China National Tobacco Corporation, Fuzhou 350003, China; 2. Zhangzhou Branch of Fujian Provincial Tobacco Corporation, Zhangzhou 363000, China; 3. School of Mathematical Sciences, Xiamen University, Xiamen 361005, China)

Abstract: [**Objective**] It employs the EM algorithm and Gaussian mixture models for dynamic data clustering, as well as focuses on 2022 sales data of tobacco brands from Zhangzhou Tobacco Company, Fujian, to uncover data patterns. [**Methods**] In this study, we use the EM algorithm with Gaussian mixture models for parameter estimation and dynamic clustering. We apply statistical software to implement the algorithm on sales data. The process involves initializing parameters, applying EM for refinement, and clustering based on probability distributions, as well as adheres to statistical principles for accurate, impartial results. [**Results**] Our findings demonstrate that the new algorithm, combining the EM algorithm with Gaussian mixture models, effectively estimates probability model parameters. This approach achieves precise multiple clustering of tobacco brands, and reveals detailed category characteristics and offering comprehensive decision support. It accurately calculates each brand's probability distribution across categories, as well as provides a robust basis for tailoring market strategies and optimizing product portfolios. The method's adherence to statistical principles ensures classification accuracy and impartiality. Recognizing the business environment's complexity, the approach allows flexible category adjustments to adapt to market changes, thus enhancing its practicality. In the study, we derive mathematical formulas for parameter updates and improve the algorithm's precision in problem-solving. Simulations and empirical analyses validate the method's superior performance with mixed distribution data and showcases its potential in real-world applications. It significantly enhances the understanding of market dynamics, and supports sales strategy formulation with data-driven insights. The research not

收稿日期:2024-07-05 **录用日期**:2024-09-06

基金项目:国家自然科学基金(12171405);闽烟司[2023]3 号(2023350000200099)

* **通信作者**:jianhualin@xmu.edu.cn

引文格式:金向阳,章惠民,王语涵,等.基于 EM 算法与混合模型的动态聚类分析[J].厦门大学学报(自然科学版),2025,64(3):727-739.

Citation: JIN X Y, ZHANG H M, WANG Y H, et al. Dynamic clustering analysis based on the EM algorithm and mixture models [J]. J Xiamen Univ Nat Sci, 2025, 64(3): 727-739. (in Chinese)

only advances fields of mixed distribution models and data classification but also aids in the digital and intelligent transformation of related industries. By introducing brand probability distributions, it adds flexibility and practicality to data analysis, thus becoming a crucial tool for enterprises to thrive in competitive markets. Finally, the method's ability to adapt to market changes provides nuanced insights and positions it as a valuable asset for businesses seeking to optimize their strategies. [**Conclusions**] This study proposes an efficient data parameter estimation and dynamic classification method based on mixture Gaussian distributions and the EM algorithm. It offers in-depth insights into brand categories and comprehensive decision support through precise data analyses. The flexibility of the method and its adaptability to market changes allow it to meet diverse client needs practically. The proposed research contributes new perspectives to mixed distribution models and data classification, and supports the digital and intelligent upgrading of industries. By emphasizing brand probability distributions, it enhances data analysis's flexibility and practicality, and becomes a key tool for businesses to excel in competitive markets.

Keywords: probability models; EM algorithm; mixture distribution; dynamic clustering

在当今竞争激烈的市场中,品牌间竞相推出多样化、个性化的产品以吸引消费者. 面对市场迅速变化和消费者偏好的不断演进,卖家面临诸多挑战. 为了应对这些挑战,准确捕捉市场动态和快速调整策略至关重要,而销售数据的深入分析和动态监控成为解决这一问题的关键. 通过精确分析销售数据,使卖家能够洞察市场需求、消费者行为和竞争对手的动态,进而能够更精准地制定战略决策,以优化产品线和营销策略. 基于数据的决策方式有助于降低市场风险,提升资源利用效率,推动可持续发展. 因此,构建一套完善的数据分析体系,对销售数据进行精确而动态的分析,已成为现代卖家提升竞争力、应对市场挑战的核心策略.

与传统意义上的聚类分析不同,本研究所使用的动态聚类分析结合了 EM 算法与混合模型,致力于探索数据背后的概率模型,从贝叶斯理论的角度上对品规的分类进行二次更新. 这种研究方法可以使销售公司依据当下的市场动态及消费者偏好及时做出品规分类等政策调整,以更好地服务社会.

在过去的几十年,前人用混合模型解决了各式各样的复杂数学建模问题,混合模型在识别模型,如信息检索、语音识别、计算机视觉、入侵检测等方面都得到了广泛应用. 混合模型还可以有效应用于机器学习中的无监督学习,在精算、市场营销、医学、生物学、天文学、工程学等领域也应用得较为深入[1],如:心理学中评估人的理解和认知能力[2],天文学中表征恒星[3],生物学中依据基因的特征数据推理系统的发生树[4],图像重建中正电子发射的断层成像问题[5].

混合 Gaussian 分布是最近几十年比较流行的工具,因为其可以近似任意连续分布且拥有良好的计算效果,因此常被用于拟合观测数据的分布. Rasmussen[6] 提出了无限 Gaussian 混合模型,克服了传统 Gaussian 混合模型中需要预先指定混合分量数量的限制.

混合模型诞生之后其参数估计是要解决的最大问题,运用最多且耳熟能详的是最大似然估计(maximum likelihood estimation, MLE),其实现目标即为找到最大化似然函数的最优解作为参数估计值. 此外,考虑到期望最大化(EM)算法[7]对初始值的选取较为敏感,初始值选择不当极其容易收敛到局部最优解或产生过拟合的现象,相对于 EM 算法贝叶斯估计通过给定参数的先验分布得到其后验分布,则较为稳定,将二者结合也是一种参数估计的解决思路.

20 世纪 70 年代,Dempster 等[7]提出 EM 算法求解模型的参数估计近似值,一时引起轰动,随之在 1984 年 Randner 等[8]介绍了在统计建模中使用混合密度的概念,并探讨了 MLE 以及 EM 算法在处理混合密度模型中的应用;2023 年,杜让[9]致力于探讨如何利用 EM 算法进行时间序列数据分析,并提出了一种基于 EM 算法的分析方法,为解决股票数据中的问题提供了理论和方法支持;在信号处理方面,Moon[10]提供了对 EM 算法的全面理解,为信号处理领域的研究者和从业者提供了重要的参考资料;Bilmes[11]提供了关于 EM 算法的简明教程以及其在参数估计中的应用,特别是在高斯混合模型和隐马尔可夫模型中的应用.

本研究的目标是提供一种新的分析工具,帮助企业更好地理解市场动态和消费者偏好的演变,从而制定更有效的市场进入和产品开发策略. 通过这项研究,我们希望能够增强行业的适应能力和竞争优势,推动数据驱动决策方面的进步.

1 理论基础

1.1 聚类分析

聚类分析是无监督机器学习中的一项关键技术,

它能根据数据间的相似性自动将数据归类,形成不同的簇群. 这种方法摆脱了对预定义标签或类别的依赖,深入挖掘数据本身的内在结构和联系,从而揭示出那些不易察觉的模式. 在诸如市场细分、用户画像描绘、异常侦测以及特征抽取等多个领域,聚类分析都发挥着不可或缺的作用. 通过聚类,可以更深入地理解数据的分布情况和特性,为后续的监督学习提供宝贵的先验知识,进而优化模型的性能和精确度.

然而,聚类分析也存在一些固有的局限性. 它对于初始参数和算法的选择极为敏感,不同的参数设置或算法选择往往会导致截然不同的聚类结果. 例如,K-means 算法要求预先指定簇的数量,这一参数的选择往往基于主观判断或需通过数据驱动的方法(如肘部法)来确定[12]. 层次聚类和基于密度的聚类在原理上存在显著差异[13],因此它们的分类结果也可能大相径庭. 即便是在使用相同算法的情况下,不同的距离指标也会导致分类结果的差异. 此外,聚类分析在处理高维度数据、形状复杂的数据集以及包含噪声或异常值的数据时,其表现可能会不尽如人意,这可能会导致错误的聚类或产生难以解读的结果.

传统的聚类方法并未深入探索待聚类数据背后的生成机制和概率原理,而只是采用了一套通用且合理的方法来进行数据分类. 然而,探索是否能用一个通用的概率模型来表征聚类数据,如 Gaussian 混合模型,不仅能描绘出数据分组的概率分布,还能通过合适的算法来估算模型参数,从而揭示出数据生成的深层逻辑[14]. 这种方法能够在提供聚类结果的同时,估计数据生成的概率模型参数,有助于更深入地理解数据的生成过程.

1.2　概率模型底层逻辑

设 $\boldsymbol{\Theta}$ 为一随机变量,其概率分布如下所示,其中第 j 个分布参数为 θ_j 的概率为 p_j:

$$\boldsymbol{\Theta} \sim \begin{Bmatrix} \theta_1 & \theta_2 \cdots & \cdots \theta_K \\ p_1 & p_2 \cdots & \cdots p_K \end{Bmatrix}.$$

将所有样本数据的产生机制总结如下:随机变量 $\boldsymbol{\Theta}$ 首先产生 $\theta_1,\cdots,\theta_K$,这里 K 为混合项数,将其作为参数向量从分布 $f_1(x \mid \theta_1),\cdots,f_K(x \mid \theta_K)$ 中分别生成 $n_1,\cdots,n_K$ 个数据(可以是高维数据),其中 $\sum_{i=1}^{K} n_i = n$.

将这 n 个数据混合在一起记作 $\boldsymbol{X}_i,\cdots,\boldsymbol{X}_n$ 作为样本,需要注意的是每一混合项是否出现未知,故将上述分布总结为形式(1),其中,$\boldsymbol{H} = (p_1,\cdots,p_K,\theta_1,\cdots,\theta_K)$ 为参数向量,p_j 为 x 来自第 j 个混合项的概率,θ_j 为其他参数向量,分布函数

$$f(x \mid \boldsymbol{H}) = \sum_{j=1}^{K} p_j f_j(x \mid \boldsymbol{\theta}_j), j = 1,\cdots,K. \quad (1)$$

本文的目的在于将这 n 个样本数据按照原来的生成机制分为 K 类,并估计出对应的模型参数 θ_j 以及混合比例参数 p_j. 下文将使用 EM 算法进行实现,并分为一维情形和高维情形进行迭代等式的推导以及算法实现的介绍.

1.3　EM 算法与一般分布结合

EM 算法是一种迭代优化算法,用来求解含有隐变量的概率模型参数的最大似然估计,在统计数据分析、机器学习等领域中有着广泛的应用,EM 算法在探寻聚类数据背后的概率模型时展现了卓越的能力. 通过其独特的迭代机制,即使在面临隐藏变量或数据缺失的复杂情境下,它也能稳健地剥离出数据的深层结构和概率分布. 这一过程不仅有助于提升对聚类数据生成机制的理解,更为后续的数据挖掘、模式识别及预测分析提供了坚实的基础. EM 算法的这一特性,使其在数据处理和机器学习领域具有不可替代的价值.

1.3.1　一维情形

设 $\boldsymbol{X} = (X_1,\cdots,X_n)$ 是一组样本,其中每个 X_i,$i = 1,\cdots,n$ 是一维的,这个样本来自于密度函数为 $f(x \mid \boldsymbol{H})$ 的混合分布总体:

$$f(x \mid \boldsymbol{H}) = \sum_{j=1}^{K} p_j f_j(x \mid \boldsymbol{\theta}_j), \quad (2)$$

其中,$\boldsymbol{H} = (p_1,\cdots,p_K,\theta_1,\cdots,\theta_K)$ 为参数向量,p_j 为 x 来自第 j 个混合项的概率,$\theta_j (j = 1,\cdots,K)$ 为总体的其他参数向量,K 为有限混合项数.

根据 EM 算法的理论,引入潜变量 $\boldsymbol{Z} = (Z_1,\cdots,Z_n)$,其中 $Z_i \in \{1,\cdots,K\}$. 当 $Z_i = j$ 时,表示样本 X_i 是由第 j 个分量产生的,即 $[X_i \mid Z_i = j] \sim f_j(x \mid \boldsymbol{\theta}_j)$,且 $\boldsymbol{Z}$ 的分布律可以假设为 $Z_i \sim P(Z_i = j) = p_j, j = 1,\cdots,K$.

因此在引入潜变量 $\boldsymbol{Z}$ 后,对数似然函数为

$$l(\boldsymbol{H} \mid \boldsymbol{X},\boldsymbol{Z}) = \sum_{i=1}^{n} \log(p_{Z_i} f_{Z_i}(X_i \mid \boldsymbol{\theta}_{Z_i})). \quad (3)$$

假设在第 $k+1$ 步迭代中估计值 $\boldsymbol{H}^{(k)}$ 已知,则通过 EM 算法的 E 步和 M 步可得迭代等式,计算出 $\boldsymbol{H}^{(k+1)}$.

E 步:

$$Q(\boldsymbol{H},\boldsymbol{H}^{(k)}) = E_Z(l(\boldsymbol{H} \mid \boldsymbol{X},\boldsymbol{Z})) =$$

$$\sum_{j=1}^{K} \sum_{i=1}^{n} [\log(p_j) \cdot p(j \mid X_i,\boldsymbol{H}^{(k)}) +$$

$$\log(f_j(X_i \mid \boldsymbol{\theta}_j)) \cdot p(j \mid X_i, \boldsymbol{H}^{(k)})], \quad (4)$$

其中第 j 个观测样本 X_i 取自于第 j 个分量的概率可表示为

$$p(Z_i = j \mid X_i, \boldsymbol{H}^{(k)}) = p(j \mid X_i, \boldsymbol{H}^{(k)}). \quad (5)$$

因此由贝叶斯公式得

$$p(j \mid X_i, \boldsymbol{H}^{(k)}) = \frac{p_j^{(k)} \cdot f_j(X_i \mid \boldsymbol{\theta}_j^{(k)})}{\sum_{j=1}^{K} p_j^{(k)} \cdot f_j(X_i \mid \boldsymbol{\theta}_j^{(k)})}. \quad (6)$$

M 步:

在 M 步中要最大化上述 $Q(\boldsymbol{H}, \boldsymbol{H}^{(k)})$,求满足期望最大化的解 $\boldsymbol{H}^{(k+1)}$,需要先求对于参数 θ 的一阶偏导数为 0 时的 $\boldsymbol{\theta}^{(k+1)}$:

$$\frac{\partial Q(\boldsymbol{H}, \boldsymbol{H}^{(k)})}{\partial \boldsymbol{\theta}} = 0. \quad (7)$$

由于 $\sum_{j=1}^{K} p_j = 1$,因此求 $p_j^{(k+1)}$ 时为带约束的优化问题,要用拉格朗日乘子法进行求解:

$$\text{Lagrange} = Q(\boldsymbol{H}, \boldsymbol{H}^{(k)}) + \lambda\Big(\sum_{j=1}^{K} p_j - 1\Big), \quad (8)$$

最终结果为

$$p_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} p(j \mid X_i, \boldsymbol{H}^{(k)}) = \frac{1}{n}\sum_{i=1}^{n} \frac{p_j^{(k)} \cdot f_j(X_i \mid \boldsymbol{\theta}_j^{(k)})}{\sum_{j=1}^{K} p_j^{(k)} \cdot f_j(X_i \mid \boldsymbol{\theta}_j^{(k)})}. \quad (9)$$

1.3.2 高维情形

在高维情形下,设 $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$ 是一组样本,其中每个 $\boldsymbol{X}_i, i = 1, \cdots, n$ 是 P 维的,这个样本来自于密度函数为 $f(x \mid \boldsymbol{H})$ 的混合分布总体:

$$f(x \mid \boldsymbol{H}) = \sum_{j=1}^{K} p_j \prod_{m=1}^{M} f_{jm}(x \mid \boldsymbol{\theta}_{jm}), \quad (10)$$

其中:$\boldsymbol{H} = (p_1, \cdots p_K, \boldsymbol{\theta}_{11}, \cdots, \boldsymbol{\theta}_{KP})$ 为参数向量;p_j 为 x 来自第 j 个混合项的概率,$j = 1, \cdots, K$,K 为有限混合项数;θ_{jm} 为总体的其他参数向量,$m = 1, \cdots, M$;M 为自变量维数.

根据 EM 算法的理论,引入潜变量 $\boldsymbol{Z} = (Z_1, \cdots, Z_n)$,其中 $Z_i \in \{1, \cdots, K\}$,当 $Z_i = j$ 时表示样本 $\boldsymbol{X}_i$ 是由第 j 个分量产生的,即 $[\boldsymbol{X}_i \mid Z_i = j] \sim \prod_{m=1}^{M} f_{jm}(X_{im} \mid \theta_{jm})$ 且 $\boldsymbol{Z}$ 的分布律可以假设为 $Z_i \sim P(Z_i = j) = p_j, j = 1, \cdots, K$.

因此在引入潜变量 $\boldsymbol{Z}$ 后,对数似然函数为

$$l(\boldsymbol{H} \mid \boldsymbol{X}, \boldsymbol{Z}) = \log\Big(\prod_{i=1}^{n} p_{Z_i} \prod_{m=1}^{M} f_{Z_i m}(X_{im} \mid \boldsymbol{\theta}_{Z_i m})\Big) = \sum_{i=1}^{n} \log\Big(p_{Z_i} \prod_{m=1}^{M} f_{Z_{im}}(X_{im} \mid \boldsymbol{\theta}_{Z_{im}})\Big). \quad (11)$$

假设在第 $k+1$ 步迭代中估计值 $\boldsymbol{H}^{(k)}$ 已知,则通过 EM 算法的 E 步和 M 步可得迭代等式,计算出 $\boldsymbol{H}^{(k+1)}$.

E 步:

$$Q(\boldsymbol{H}, \boldsymbol{H}^{(k)}) = E_Z(l(\boldsymbol{H} \mid \boldsymbol{X}, \boldsymbol{Z})) = \sum_{j=1}^{K}\sum_{i=1}^{n}[\log(p_j) \cdot p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(k)}) + \sum_{m=1}^{M} \log(f_{jm}(X_{im} \mid \boldsymbol{\theta}_{jm})) \cdot p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(k)})], \quad (12)$$

其中第 j 个观测样本 $\boldsymbol{X}_i$ 取自于第 j 个分量的概率可表示为

$$p(Z_i = j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(k)}) = p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(k)}), \quad (13)$$

因此由贝叶斯公式得:

$$p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(k)}) = \frac{p_j^{(k)} \cdot f_j(\boldsymbol{X}_i \mid \boldsymbol{\theta}_j^{(k)})}{\sum_{j=1}^{K} p_j^{(k)} \cdot f_j(\boldsymbol{X}_i \mid \boldsymbol{\theta}_j^{(k)})} = \frac{p_j^{(k)} \cdot \prod_{m=1}^{M} f_{jm}(X_{im} \mid \boldsymbol{\theta}_{jm}^{(k)})}{\sum_{j=1}^{K} p_j^{(k)} \cdot \prod_{m=1}^{M} f_{jm}(X_{im} \mid \boldsymbol{\theta}_{jm}^{(k)})}. \quad (14)$$

M 步:

在 M 步中要最大化上述 $Q(\boldsymbol{H}, \boldsymbol{H}^{(k)})$,求满足期望最大化的解 $\boldsymbol{H}^{(k+1)}$,需要先求参数 θ 的一阶偏导数为 0 时的 $\boldsymbol{\theta}^{(k+1)}$:

$$\frac{\partial Q(\boldsymbol{H}, \boldsymbol{H}^{(k)})}{\partial \boldsymbol{\theta}} = 0. \quad (15)$$

由于 $\sum_{j=1}^{K} p_j = 1$,因此求 $p_j^{(k+1)}$ 时为带约束的优化问题,要用拉格朗日乘子法进行求解:

$$\text{Lagrange} = Q(\boldsymbol{H}, \boldsymbol{H}^{(k)}) + \lambda\Big(\sum_{j=1}^{K} p_j - 1\Big), \quad (16)$$

最终结果为

$$p_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(k)}) = \frac{1}{n}\sum_{i=1}^{n} \frac{p_j^{(k)} \cdot \prod_{m=1}^{M} f_{jm}(X_{im} \mid \boldsymbol{\theta}_{jm}^{(k)})}{\sum_{j=1}^{K} p_j^{(k)} \cdot \prod_{m=1}^{M} f_{jm}(X_{im} \mid \boldsymbol{\theta}_{jm}^{(k)})}. \quad (17)$$

由于本节为一般分布情形,没有具体表达式,因此关于参数可识别的问题请见 1.4 节.

1.4 EM 算法与混合 Gaussian 分布结合

混合 Gaussian 分布因其灵活性、表示能力、泛化能力、聚类能力和生成数据的能力而备受推崇. 其灵活性使其能够适应各种形状的数据分布,出色的泛化能力使其在处理不确定性和异常值时表现鲁棒,而聚类能力使其成为强大的聚类算法. 同时,混合 Gaussian 分布还可以生成与原始数据相似的新数据样本,为数据分析和建模提供了便利. 因此将 1.3 节的一般算法应用于混合 Gaussian 分布,在本节中给出一维和高维情形下算法的迭代等式的推导过程以及算法的具体流程.

1.4.1　一维情形

设 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 是一组样本,其中每个 X_i, $i=1,\cdots,n$ 是一维的,这个样本来自于密度函数为 $f(x\mid\boldsymbol{H})$ 的混合高斯分布总体:

$$f(x\mid\boldsymbol{H})=\sum_{j=1}^{K}p_jf_j(x\mid\boldsymbol{\theta}_j)=\sum_{j=1}^{K}p_j\frac{1}{\sqrt{2\pi}\sigma_j}\mathrm{e}^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}},$$

其中 μ_j 为第 j 个混合项的期望,σ_j^2 为第 j 个混合项的方差,记 $\boldsymbol{\theta}_j=(\mu_j,\sigma_j^2)$, $j=1,\cdots,K$, 则 $\boldsymbol{H}=(p_1,\cdots,p_K,\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_K)$ 为参数向量,p_j 为 x 来自第 j 个混合项的概率,K 为有限混合项数.

如一般情形的思想将总体密度函数代入计算迭代算式,省略中间步骤,现将该算法的步骤总结如下:

步骤 0　选用合适的方法确定初始值,记为 $\mu_j^{(0)}$, $\sigma_j^{(0)}$, $p_j^{(0)}$, 其中 $j=1,\cdots,K$.

步骤 1　利用迭代等式首先确定 $p(j\mid X_i,\boldsymbol{H}^{(0)})$:

$$p(j\mid X_i,\boldsymbol{H}^{(0)})=\frac{p_j^{(0)}\cdot\frac{1}{\sqrt{2\pi}\sigma_j^{(0)}}\mathrm{e}^{-\frac{(x_i-\mu_j^{(0)})^2}{2\sigma_j^{2(0)}}}}{\sum_{j=1}^{K}p_j^{(0)}\cdot\frac{1}{\sqrt{2\pi}\sigma_j^{(0)}}\mathrm{e}^{-\frac{(x_i-\mu_j^{(0)})^2}{2\sigma_j^{2(0)}}}}.\tag{18}$$

步骤 2　然后利用迭代等式确定 $\mu_j^{(1)}$, $\sigma_j^{2(1)}$, $p_j^{(1)}$:

$$\mu_j^{(1)}=\frac{\sum_{i=1}^{n}X_i\cdot p(j\mid X_i,\boldsymbol{H}^{(0)})}{\sum_{i=1}^{n}p(j\mid X_i,\boldsymbol{H}^{(0)})},\tag{19}$$

$$\sigma_j^{2(1)}=\frac{\sum_{i=1}^{n}(X_i-\mu_j^{(1)})^2\cdot p(j\mid X_i,\boldsymbol{H}^{(0)})}{\sum_{i=1}^{n}p(j\mid X_i,\boldsymbol{H}^{(0)})},\tag{20}$$

$$p_j^{(1)}=\frac{1}{n}\sum_{i=1}^{n}p(j\mid X_i,\boldsymbol{H}^{(0)}).\tag{21}$$

步骤 3　然后利用更新后的参数确定 $p(j\mid\boldsymbol{X}_i,\boldsymbol{H}^{(1)})$:

$$p(j\mid X_i,\boldsymbol{H}^{(1)})=\frac{p_j^{(1)}\cdot\frac{1}{\sqrt{2\pi}\sigma_j^{(1)}}\mathrm{e}^{-\frac{(X_i-\mu_j^{(1)})^2}{2\sigma_j^{2(1)}}}}{\sum_{j=1}^{K}p_j^{(1)}\cdot\frac{1}{\sqrt{2\pi}\sigma_j^{(1)}}\mathrm{e}^{-\frac{(X_i-\mu_j^{(1)})^2}{2\sigma_j^{2(1)}}}}.\tag{22}$$

步骤 4　重复上述步骤直到达到最大迭代步数或者参数全部收敛.

最后,利用 Gaussian 分布一维情形的表达式进行模型可识别的说明,高维情形可类比说明. 证明模型可识别的思路为首先求出 Q 表达式关于参数 μ 和 σ 的信息矩阵,再证明该矩阵的正定性. 设信息矩阵为 $\boldsymbol{I}$,有如下表达式:

$$\boldsymbol{I}=\begin{bmatrix}a_{1,1}&\cdots&a_{1,2K}\\ \cdots&\cdots&\cdots\\ a_{2K,1}&\cdots&a_{2K,2K}\end{bmatrix}.\tag{23}$$

经计算可得:

$$a_{j,j}=\frac{\partial^2Q}{\partial\mu_j^2}=-\frac{\sum_{i=1}^{n}p(j\mid X_i,\boldsymbol{H}^{(k)})}{\sigma_j^2},1\leqslant j\leqslant K,\tag{24}$$

$$a_{j,j}=\frac{\partial^2Q}{\partial\theta_j^2}=\sum_{i=1}^{n}\frac{1}{2}p(j\mid X_i,\boldsymbol{H}^{(k)})\sigma_j^{-4}-\sum_{i=1}^{n}(X_i-\mu_j)^2p(j\mid X_i,\boldsymbol{H}^{(k)})\sigma_j^{-6},\quad K+1\leqslant j\leqslant 2K,\tag{25}$$

$$a_{j,j+K}=\frac{\partial^2Q}{\partial\mu_j\partial\theta_j}=-\frac{\sum_{i=1}^{n}(X_i-\mu_j)p(j\mid X_i,\boldsymbol{H}^{(k)})}{\sigma_j^4},\quad 1\leqslant j\leqslant K,\tag{26}$$

$$a_{j_1,j_2}=0,\text{ 其他元素均为 }0.\tag{27}$$

利用矩阵行列式可得 $\det(\boldsymbol{I})=a_{1,1}\cdot a_{2,2}\cdot\cdots\cdot a_{2K,2K}$, 由上式易得 $a_{j,j}\neq 0, j=1,\cdots,2K$, 且 $a_{j,j}\cdot a_{j,j+K}>0, 1\leqslant j\leqslant K$, 因此可得 $\det(\boldsymbol{I})>0$, 故信息矩阵正定,模型可识别. 由于模拟实验以及实证分析均为销售数据,数据量不大,因此对参数估计的收敛速度影响不大,所以没有考虑.

1.4.2　高维情形

在高维情形下,设 $\boldsymbol{X}=(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)$ 是一组样本,其中每个 $\boldsymbol{X}_i(i=1,\cdots,n)$ 是 P 维的,这个样本来自于密度函数为 $f(\boldsymbol{x}\mid\boldsymbol{H})$ 的高斯混合分布总体:

$$f(\boldsymbol{x}\mid\boldsymbol{H})=\sum_{j=1}^{K}p_j\prod_{m=1}^{M}f_{jm}(\boldsymbol{x}\mid\boldsymbol{\theta}_{jm})=\sum_{j=1}^{K}p_j\prod_{m=1}^{M}\frac{1}{\sqrt{2\pi}\sigma_{jm}}\mathrm{e}^{-\frac{(x-\mu_{jm})^2}{2\sigma_{jm}^2}},\tag{28}$$

其中:μ_{jm} 为第 j 个混合项的第 m 个维度上的期望;σ_{jm}^2 为第 j 个混合项的第 m 个维度上的方差;记 $\boldsymbol{\theta}_{jm}=(\mu_{jm},\sigma_{jm}^2)$, $j=1,\cdots,K, m=1,\cdots,M$, 则 $\boldsymbol{H}=(p_1,\cdots,p_K,\boldsymbol{\theta}_{11},\cdots,\boldsymbol{\theta}_{KM})$ 为参数向量,p_j 为自变量来自第 j 个混合项的概率,K 为有限混合项数,M 为自变量维度.

类似一维情形,将该算法的步骤总结如下:

步骤 0　选用合适的方法确定初始值,记为 $\mu_{jm}^{(0)}$, $\sigma_{jm}^{2(0)}$, $p_j^{(0)}$,其中 $j=1,\cdots,K; m=1,\cdots,M$.

步骤 1　利用迭代等式首先确定 $p(j\mid\boldsymbol{X}_i,\boldsymbol{H}^{(0)})$:

$p(j\mid\boldsymbol{X}_i,\boldsymbol{H}^{(0)})=$

$$\frac{p_j^{(0)} \cdot \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_{jm}^{(0)}} \mathrm{e}^{-\frac{(X_{im}-\mu_{jm}^{(0)})^2}{2\sigma_{jm}^{2(0)}}}}{\sum_{j=1}^{K} p_j^{(0)} \cdot \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_{jm}^{(0)}} \mathrm{e}^{-\frac{(X_{im}-\mu_{jm}^{(0)})^2}{2\sigma_{jm}^{2(0)}}}}. \quad (29)$$

步骤 2　然后利用迭代等式确定 $\mu_{jm}^{(1)}$，$\sigma_{jm}^{2(1)}$，$p_j^{(1)}$：

$$\mu_{jm}^{(1)} = \frac{\sum_{i=1}^{n} X_{im} \cdot p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(0)})}{\sum_{i=1}^{n} p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(0)})}, \quad (30)$$

$$\sigma_{jm}^{2(1)} = \frac{\sum_{i=1}^{n} (X_{im} - \mu_{jm}^{(1)})^2 \cdot p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(0)})}{\sum_{i=1}^{n} p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(0)})}, \quad (31)$$

$$p_j^{(1)} = \frac{1}{n} \sum_{i=1}^{n} p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(0)}). \quad (32)$$

步骤 3　然后利用更新后的参数确定 $p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(1)})$：

$$p(j \mid \boldsymbol{X}_i, \boldsymbol{H}^{(1)}) = \frac{p_j^{(1)} \cdot \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_{jm}^{(1)}} \mathrm{e}^{-\frac{(X_{im}-\mu_{jm}^{(1)})^2}{2\sigma_{jm}^{2(1)}}}}{\sum_{j=1}^{K} p_j^{(1)} \cdot \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_{jm}^{(1)}} \mathrm{e}^{-\frac{(X_{im}-\mu_{jm}^{(1)})^2}{2\sigma_{jm}^{2(1)}}}}. \quad (33)$$

步骤 4　重复上述步骤直到达到最大迭代步数或者参数全部收敛.

2　模拟实验

模拟实验中，设计混合 Gaussian 分布的一维和高维情形.

一维情形依据 $\mu_1 = 0.000\ 0$，$\mu_2 = 5.000\ 0$，$\mu_3 = 10.000\ 0$，$\mu_4 = 15.000\ 0$，$\mu_5 = 20.000\ 0$；$\sigma_1^2 = 1.000\ 0$，$\sigma_2^2 = 2.250\ 0$，$\sigma_3^2 = 4.000\ 0$，$\sigma_4^2 = 2.250\ 0$，$\sigma_5^2 = 1.000\ 0$；$p_1 = 0.200\ 0$，$p_2 = 0.200\ 0$，$p_3 = 0.200\ 0$，$p_4 = 0.200\ 0$，$p_5 = 0.200\ 0$ 的参数初始设置生成 100 个样本，参数估计的均方误差为 $E_\mu^{\mathrm{MS}} = 0.022\ 1$、$E_\sigma^{\mathrm{MS}} = 0.007\ 9$、$E_p^{\mathrm{MS}} = 0.000\ 1$. 然后利用模拟生成的样本进行动态聚类，第一次聚类结果如表 1 所示，经过第一次聚类后将五类中的数据更新，第二次聚类结果如表 2 所示，聚类结果的混淆矩阵如表 3 所示.

高维情形依据以下参数初始设置生成 100×10 个样本：

$\mu_1 = (1.000\ 0, 6.000\ 0, 11.000\ 0, 16.000\ 0, 21.000\ 0, 26.000\ 0, 31.000\ 0, 36.000\ 0, 41.000\ 0, 46.000\ 0)$，

$\mu_2 = (2.000\ 0, 7.000\ 0, 12.000\ 0, 17.000\ 0, 22.000\ 0, 27.000\ 0, 32.000\ 0, 37.000\ 0, 42.000\ 0, 47.000\ 0)$，

$\mu_3 = (3.000\ 0, 8.000\ 0, 13.000\ 0, 18.000\ 0, 23.000\ 0, 28.000\ 0, 33.000\ 0, 38.000\ 0, 43.000\ 0, 48.000\ 0)$，

$\mu_4 = (4.000\ 0, 9.000\ 0, 14.000\ 0, 19.000\ 0, 24.000\ 0, 29.000\ 0, 34.000\ 0, 39.000\ 0, 44.000\ 0, 49.000\ 0)$，

表 1　混合高斯分布一维动态聚类结果

Tab. 1　One dimensional dynamic clustering results of mixed Gaussian distribution

初始分类	聚类结果																				
P1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
P2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
P3	3	3	3	4	3	3	3	3	3	3	3	3	3	4	3	4	2	3	3	3	
P4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	
P5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

表 2　混合高斯分布一维二次动态聚类结果

Tab. 2　One dimensional quadratic dynamic clustering results of mixed Gaussian distribution

初始分类	聚类结果																					
P1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
P2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
P3	3	3	3	3	3	3	3	3	3	3	3	3	4	3	3	3	3					
P4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	3	3
P5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5		

表 3　高斯分布一维情形动态聚类混淆矩阵

Tab. 3　Confusion matrix of Gaussian distribution one-dimensional dynamic clustering

R	P				
	I	II	III	IV	V
I	20	0	0	0	0
II	0	20	0	0	0
III	0	1	16	3	0
IV	0	0	1	19	0
V	0	0	0	0	20

$\mu_5 = (5.000\ 0, 10.000\ 0, 15.000\ 0, 20.000\ 0, 25.000\ 0, 30.000\ 0, 35.000\ 0, 40.000\ 0, 45.000\ 0, 50.000\ 0)$,

$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = (1.000\ 0, 1.000\ 0, 1.000\ 0, 1.000\ 0, 1.000\ 0, 1.000\ 0, 1.000\ 0, 1.000\ 0, 1.000\ 0, 1.000\ 0)$,

$p_1 = 0.200\ 0, p_2 = 0.200\ 0, p_3 = 0.200\ 0, p_4 = 0.200\ 0, p_5 = 0.200\ 0$. 参数估计的均方误差为 $E_\mu^{MS} = 0.011\ 5$、$E_\sigma^{MS} = 0.175\ 7$、E_p^{MS} 约为 0. 然后利用模拟生成的样本进行动态聚类,第一次聚类结果如表 4 所示,经过第一次聚类后将五类中的数据更新,第二次聚类结果如表 5 所示,聚类结果的混淆矩阵如表 6 所示.

由于每类样本数相同且每类样本同等重要,因此选用宏平均(Macro-Averaged)来计算各种指标,从聚类效果表 7 可知,Accuracy、Macro-Averaged Precision 以及 Macro- Averaged F_1 Score 都接近 1,可以说明本文的聚类方法具有可行性以及较高的准确率.

表 4　混合高斯分布高维动态聚类结果

Tab. 4　High-dimensional dynamic clustering results of mixed Gaussian distribution

初始分类	聚类结果																			
P1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
P2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2
P3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3	3
P4	3	4	4	3	4	4	4	3	3	4	4	5	4	4	4	4	4	4	3	4
P5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

表 5　混合高斯分布高维二次动态聚类结果

Tab. 5　High-dimensional quadratic olynamic clustering results of mixed Gaussian distribution

初始分类	聚类结果																								
P1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
P2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2					
P3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3
P4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4										
P5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5					

表 6　高斯分布高维情形动态聚类混淆矩阵

Tab. 6　Confusion matrix of Gaussian distribution high-dimensional dynamic clustering

R	P				
	I	II	III	IV	V
I	20	0	0	0	0
II	0	19	1	0	0
III	0	1	19	0	0
IV	0	0	5	14	1
V	0	0	0	1	19

3　实证分析

3.1　数据说明和预处理

数据预处理是任何数据分析项目成功的关键步骤,特别是在处理实际业务数据时,这一步骤显得尤为重要. 本研究中的数据集由漳州烟草公司提供,涵盖了 2019 至 2021 年间的销售记录,包括但不限于销售量、销售额、动销率、社会存销比等多维度信息. 在进行深入分析之前,对原始数据执行一系列预处理步

表 7 聚类结果

Tab. 7 Clustering results

维度	Accuracy	Macro-Averaged Precision	Macro-Averaged F_1 Score
一维情形	0.950 0	0.988 2	0.972 4
高维情形	0.910 0	0.955 7	0.916 0

骤,以确保数据质量和分析的准确性.

首先,对数据集进行清洗,包括去除重复记录、修正不一致的数据格式和处理缺失值. 特别地,对于缺失值的处理,采取不同策略:对于少量缺失的数值型数据,使用该字段的中位数进行填充;对于分类数据,使用最频繁出现的类别进行替代. 这一步骤旨在保留尽可能多的数据,同时减少缺失值对模型性能的潜在影响.

其次,针对高维数据特征,通过分析特征间的相关性和重要性去除一些冗余和不相关的特征,同时减少模型的复杂度.

经数据预处理,最终得到 156 个样本的 12 个特征的数据,数据特征分别为:销量、条均价、品类销售份额、品类销售额增长率、品牌认知度、总客户数、上柜率、单品溢价率、零售价格指数、零售毛利率、销售速率.

这里并未提前进行数据的标准化,这是由于下文要针对不同模型的特征采用不同的方法进行数据的标准化.

在应用本文的模型进行实验之前,由于并未指定初始聚类数,因此首先使用肘部法、轮廓图等常见方法,确定聚成五类最为合理(也可以根据实际需求提前指定聚类数量),然后使用 K 均值、层次聚类、密度聚类等基础分类算法进行聚类,以此为对照组以及基础来进行下述模型的实验,这五类品规的特征总结如下.

N_1:高潜力品规,具有较强的发展及培育潜力,消费认可度高,当前市场需求较强且供给缺口较大. 如中华(双中支)、七匹狼(软灰)等.

N_2:低潜力品规,具有一定的发展和培育潜力,当前市场需求略大于供给,供给缺口不大. 如七匹狼(古田红车灰)、七匹狼(尚品)等.

N_3:平衡型品规,供需处于相对平衡状态,增量空间较小. 如七匹狼(古田金中支)、云烟(软如意).

N_4:弱需求品规,有一部分消费群体,产生较小的市场需求,有需求但终端动销速度较慢. 如冬虫夏草(和润)、七匹狼(厦门)等.

N_5:待转化品规,供需处于严重的失衡状态,基本无消费需求,大量库存积压在终端门店. 如天子(千里江山细支)、泰山(儒风细支)等.

3.2 实验目的说明

首先,以上述已经分好类的数据为基础,利用其类样本均值、类样本方差、似然函数等数字特征为依据产生的参数作为初始值,使用前述算法迭代计算每个模型的参数,可以观察与原始类的参数差距,衡量本文算法与上述分类算法的评价标准相似性,但这不能说明每种分类算法的优劣.

其次,对新数据实现动态二次聚类,这里用到的思想是贝叶斯定理:

$$P(j \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid j)P(j)}{P(\boldsymbol{X})} = \frac{P(\boldsymbol{X} \mid j)P(j)}{\sum_{j=1}^{K} P(\boldsymbol{X} \mid j)P(j)}, \tag{34}$$

其中:$P(j|\boldsymbol{X})$表示新数据 $\boldsymbol{X}$ 被分在第 j 类中的概率,是后验概率;$P(\boldsymbol{X}|j)$表示新数据 $\boldsymbol{X}$ 就在第 j 类中的概率,相当于似然;$P(j)$就是上文中的 p_j,称为先验概率;K 表示类别数,在这里为 5.

利用估计好参数的模型计算上述后验概率 $P(j|\boldsymbol{X})$,$j=1,\cdots,K$,并比较,选取最大的 $P(j|\boldsymbol{X})$对应的类别 j,依此可以实现对新样本的聚类.

3.3 混合 Gaussian 分布参数估计

3.3.1 一维情形

在一维情形中仅选取“销量”一列来进行分析,即数据维度为 156×1,μ 和 σ^2 的初始值设置为每一类数据的样本均值和样本方差,使用 $X_i = \dfrac{X_i - \overline{X}}{S(X)}$ 进行数据预处理,$\overline{X}$ 为样本平均值,$S(X)$为样本标准差. 使用混合 Gaussian 分布对应的算法,设置迭代次数为 100,得到参数估计结果为 $\mu_1 = 0.000\ 0$,$\mu_2 = 0.000\ 0$,$\mu_3 = 0.000\ 0$,$\mu_4 = 0.000\ 0$,$\mu_5 = 0.000\ 0$;$\sigma_1^2 = 0.967\ 9$,$\sigma_2^2 = 0.967\ 9$,$\sigma_3^2 = 0.967\ 9$,$\sigma_4^2 = 0.967\ 9$,$\sigma_5^2 = 0.967\ 9$;$p_1 = 0.241\ 9$,$p_2 = 0.126\ 9$,$p_3 = 0.243\ 3$,$p_4 = 0.233\ 1$,$p_5 = 0.136\ 5$.

3.3.2 高维情形

在高维情形中选取所有列进行分析,即数据维度为 156×12,μ 和 σ^2 的初始值设置为每一类数据的每一特征的样本均值和样本方差,p 的初始值为原始五类占总数的比例. 得到参数估计结果为:

$$\mu_1 = (-0.058\ 2, 0.008\ 6, -0.003\ 3, -0.047\ 8, 0.005\ 8, -0.053\ 8, -0.016\ 2, -0.005\ 4,$$

0.001 2,0.008 2,−0.033 6),

$\mu_{\cdot 2}$ = (−0.085 7,0.012 7,−0.004 0,−0.014 3,
−0.070 8,0.011 9,−0.082 9,−0.029 9,
−0.010 9,0.000 2,0.010 1,−0.044 7),

$\mu_{\cdot 3}$ = (−0.102 5,0.019 1,−0.007 8,−0.019 9,
−0.086 7,0.017 3,−0.102 4,−0.045 0,
−0.017 1,−0.003 0,0.011 7,−0.049 3),

$\mu_{\cdot 4}$ = (−0.128 3,0.123 5,−0.054 1,−0.130 6,
−0.096 6,0.008 2,−0.136 6,−0.047 8,
0.029 3,0.020 6,0.085 7,−0.014 8),

$\mu_{\cdot 5}$ = (0.605 5,−0.230 0,0.092 7,0.245 2,
0.501 1,−0.082 3,0.620 0,0.233 3,0.021 3,
−0.022 1,−0.162 0,0.229 4);

$\sigma^2_{\cdot 1}$ = (0.799 6,0.963 6,0.957 5,0.959 4,0.831 6,
0.971 7,0.836 0,0.975 4,0.972 4,0.975 9,
0.965 8,0.877 8),

$\sigma^2_{\cdot 2}$ = (0.743 0,0.955 2,0.945 8,0.947 7,0.790 1,
0.979 4,0.785 2,0.980 5,0.966 9,0.975 2,
0.961 1,0.856 8),

$\sigma^2_{\cdot 3}$ = (0.716 2,0.956 5,0.937 8,0.945 9,0.764 6,
0.987 4,0.758 7,0.986 4,0.960 2,0.972 1,
0.958 0,0.851 1),

$\sigma^2_{\cdot 4}$ = (0.679 2,1.137 1,0.954 5,1.120 5,0.729 8,
0.988 3,0.723 1,0.988 6,1.150 7,1.110 2,
1.084 7,0.891 1),

$\sigma^2_{\cdot 5}$ = (1.969 5,0.755 5,1.113 9,0.802 9,1.843 9,
0.857 4,1.723 7,0.808 5,0.743 7,0.748 3,
0.813 6,1.496 6);

p_1 = 0.238 8, p_2 = 0.124 4, p_3 = 0.239 2, p_4 = 0.237 7, p_5 = 0.155 4.

3.4　动态聚类

在以下对于数据聚类的实证分析中,继续应用混合 Gaussian 分布对应的算法作为主要的分析工具,利用前述 EM 算法迭代优化参数.

3.4.1　一维情形

步骤 0　使用我们的数据训练模型,得到估计的参数 μ_j 和 σ_j;

步骤 1　首先随机在数据中选取一个作为新数据,这里仅仅使用“销量”这一特征,记为数据 x;

步骤 2　计算原始数据“销量”一列的均值 μ 以及标准差 σ;

步骤 3　将新数据 x 标准化,即:

$$x = \frac{x-\mu}{\sigma};$$

步骤 4　将标准化后的数据代入以下公式进行计算后验概率:

$$P(j \mid x) = \frac{p_j \dfrac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}}{\sum_{j=1}^{K} p_j \dfrac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}}; \tag{35}$$

步骤 5　比较 5 个后验概率的大小,选择最大的后验概率对应的 j 作为其最后分类.

以下为一维情形一个数据的编程实现结果:第一个数据 586.21(初始分类为第一类),计算各类的后验概率分别为:0.246 4、0.129 2、0.247 8、0.237 5、0.139 1,可以看出最大的为第三类对应的后验概率,其次为第一类,且两者相差不大,说明这个数据依据特征“销量”应归入第三类或第一类.对于所有数据的二次动态聚类结果如表 8 所示.

从表 8 可以看出被分在第一类、第三类、第五类的数据点比较少,因此可以认为这三类不是当时市场的主流和群众的消费主流.这里我们去除这三类数据点,仅留下第二类和第四类进行再聚类,结果如表 9 所示.

去除一些样本点后,可以发现被分到第五类的样本居多,被分到第一、二、三、四类的样本较少,因

表 8　烟草公司销售数据基于混合高斯分布的一维二次动态聚类结果

Tab.8　One-dimensional quadratic dynamic clustering results of tobacco company sales data based on mixed Gaussian distribution

初始分类	聚类结果
P1	4 2 2 2 2 4 4 1 1 1 2 4 2 4 2 3 4 2 4 4 4 4 4 4 2 3 4 4
P2	4 4 4 4 5 4 2 5 4 4 4 4 4 4 4 4 4 4
P3	4 2 3 2 2 4 2 2 3 1 4 4 5 3 2 3 2 2 3 5 2 1 3 4 4 4 4 5 4 2 5 2 2 2 4 3 4 4
P4	4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 4 4 4 4 4 4 2 4 4 4 4 2 4 4 4 4 4 4 4 4 4 4 4
P5	4 4 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

表 9　烟草公司销售数据基于混合高斯分布的一维二次动态聚类结果(去除部分样本点)

Tab. 9　One-dimensional quadratic dynamic clustering results of tobacco company sales data based on mixed Gaussian distribution (excluding some data points)

初始分类	聚类结果
P1	1 1 1 1 3 4 4 1 1 3 1 1 4 3 4 4 4 4 4 5 1 5 4
P2	5 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5
P3	5 3 3 1 5 3 3 5 5 3 1 3 1 4 4 1 4 5 3 1 1 3 4 4 4
P4	5 5 5 5 5 5 4 5 5 5 5 5 5 1 4 5 5 4 5 1 4 5 1 4 5 5 5 1 5 5 5 4 5 4 5 4 5 5 5 5 5 5 5
P5	5 4 4 5 5 5 5 5 5 5 5 4 4 5 5 5 5 5 5 5 5 5 5 5 4 4 5 4

此这里可以断定分成五类也许不合适，而两类更合适，因此再将上述样本数据分成两类，将表 8 中第四类作为新的第一类初始分类，共有 109 个样本点，将第二类作为新的第二类初始分类，共有 26 个样本点，最终结果如表 10 所示，可以看到分为两类此时是合适的.

表 10　烟草公司销售数据基于混合高斯分布的一维二次动态聚类结果(分为两类)

Tab. 10　One-dimensional quadratic dynamic clustering results of tobacco company sales data based on mixed Gaussian distribution (divided into two categories)

初始分类	聚类结果
P1	1 1
P2	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

3.4.2　高维情形

步骤 0　使用数据训练模型，得到估计的参数 μ_{jm} 和 σ_{jm}；

步骤 1　首先随机在我们的数据中选取一个作为新数据，记为数据 x；

步骤 2　计算原始数据每一列的均值 μ_m 以及标准差 σ_m；

步骤 3　将新数据 x 按列标准化，即：

$$x_m=\frac{x_m-\mu_m}{\sigma_m};$$

步骤 4　将标准化后的数据代入以下公式进行计算后验概率：

$$P(j\mid x)=\frac{p_j\prod_{m=1}^{M}\frac{1}{\sqrt{2\pi}\sigma_{jm}}\mathrm{e}^{-\frac{(x-\mu_{jm})^2}{2\sigma_{jm}^2}}}{\sum_{j=1}^{K}p_j\prod_{m=1}^{M}\frac{1}{\sqrt{2\pi}\sigma_{jm}}\mathrm{e}^{-\frac{(x-\mu_{jm})^2}{2\sigma_{jm}^2}}};\quad(36)$$

步骤 5　比较 5 个后验概率的大小，选择最大的后验概率对应的 j 作为其最后分类.

以下为高维情形一个数据的编程实现结果：选取第一个数据(586.21，121.90，0.23，0.06，52 991 790.3，0.34，177 101，0.75，1.05，1.00，0.13，0.03)(初始分类为第一类)，计算各类的后验概率分别为：0.179 5、0.097 8、0.190 7、0.275 6、0.185 9 可以看出最大的为第四类对应的后验概率，说明该数据依据这 12 个特征应归入第四类. 对于所有数据的二次动态聚类结果如表 11 所示.

从表 11 可以看出被重新分类到第一类的样本点极少，而且没有被分到第二类中的数据点，因此删去这次分类中被分到第一类的数据，继续将剩下的样本点分为五类，所得结果如表 12 所示.

从表 12 可以看到，只有被分到第一类、第四类、第五类的数据，第二类和第三类没有被分到数据，因此考虑将数据分为三类，与一维相同的想法，将第一次结果中被分到第三类的数据作为新的初始第一类，被分到第四类的数据作为新的初始第二类，被分到第五类的数据作为新的初始第三类，最终分成三类的结果如表 13，可以看到最终数据均被分到了第一类，说明用全部指标作为评判标准所有品规在我们的算法

表 11　烟草公司销售数据基于混合高斯分布的高维二次动态聚类结果

Tab. 11　High-dimensional quadratic dynamic clustering results of tobacco company sales data based on mixed Gaussian distribution

初始分类	聚类结果
P1	3 4 3 4 3 3 3 5 5 5 4 3 3 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3
P2	3 3 3 5 5 3 5 5 3 3 3 4 3 3 3 3 3 4
P3	3 3 5 3 3 5 3 4 5 5 4 4 5 5 3 5 1 4 5 1 1 5 3 3 3 3 3 3 4 3 3 3 3 4 3 3 4 3
P4	3 4 3 3 4 3 3 3 3 3 3 3 3 5 5 3 3 3 3 5 1 3 5 4 3 4 3 5 4 3 3 4 3 1 4 3 3 3 4 3 4 4 4
P5	3 4 5 5 3 4 4 3 3 3 4 4 3 3 3 3 3 3 4 4 3 4 4 3 3 1 3 3 4

表 12　烟草公司销售数据基于混合高斯分布的高维二次动态聚类结果（去除部分样本点）

Tab. 12　High-dimensional quadratic dynamic clustering results of tobacco company sales data based on mixed Gaussian distribution (excluding some data points)

初始分类	聚类结果
P1	1 4 5 4 5 1 1 4 4 5 1 1 1 5 1 4 1 1 1 1 5 1 1
P2	1 4 1 5 5 5 4 4 4 4 1 1 4 1 4 4
P3	1 5 5 1 4 5 4 4 4 5 5 4 5 1 1 1 1 4 5 5 5 5 1 4 1
P4	1 4 1 1 4 1 1 1 1 4 1 1 4 5 5 1 1 5 1 5 5 1 5 4 4 4 4 5 4 4 1 4 4 5 4 1
P5	1 4 5 1 4 4 1 1 4 4 4 5 5 4 1 1 1 4 4 1 4 4 5 5 5 5 5 4

表 13　烟草公司销售数据基于混合高斯分布的高维二次动态聚类结果（分成三类）

Tab. 13　High-dimensional quadratic dynamic clustering results of tobacco company sales data based on mixed Gaussian distribution (divided into three categories)

初始分类	聚类结果
P1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
P2	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
P3	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

中可以视为有相同的发展潜力.

要说明的是，分类结果看似是“静态”，实则可以看作“动态”，其背后的逻辑其实是对于每一个样本点，都会得到一个概率分布，因此两种情形都会得到 109 个概率分布，最终所呈现的分类结果为在这个分布中最大的 p_j 对应的那一类 j.

这里使用保险精算中的布尔曼信度对一维分为两类的现象进行解释. 布尔曼信度因子 Z 是用于确定先验信息和个体经验数据在预测或分类模型中的相对重要性的一个权重，通常表示为：

$$Z = \frac{T}{T+K},$$

$$K = v/a,$$

$$v = E[V[X \mid \theta]],$$

$$a = V[E[X \mid \theta]],$$

其中：T 为经验数据的观察次数（如年数），K 为与过程方差和假设均值方差相关的参数，V 表示求方差，v 为组内差异，a 为组间差异.

在这里计算一维情形中的布尔曼信度以对结果进行说明，经计算可得：

$$E[X_1 \mid \theta_1] = 192.39,$$

$$E[X_2 \mid \theta_2] = 1\,153.94,$$

$$V[X_1 \mid \theta_1] = 32\,250.31,$$

$$V[X_2 \mid \theta_2] = 136\,194.76,$$

因此，有

$$E[X\mid\theta]\sim\begin{Bmatrix}E[X_1\mid\theta_1] & E[X_2\mid\theta_2]\\ p_1 & p_2\end{Bmatrix}=\begin{Bmatrix}192.39 & 1\,153.94\\ \frac{109}{135} & \frac{26}{135}\end{Bmatrix},$$

$$V[X\mid\theta]\sim\begin{Bmatrix}Var[X_1\mid\theta_1] & Var[X_2\mid\theta_2]\\ p_1 & p_2\end{Bmatrix}=\begin{Bmatrix}32\,250.31 & 136\,194.76\\ \frac{109}{135} & \frac{26}{135}\end{Bmatrix}.$$

故

$$v=E[V[X\mid\theta]]=32\,250.31\times\frac{109}{135}+136\,194.76\times\frac{26}{135}=52\,269.24,$$

$$a=V[E[X\mid\theta]]=\frac{109}{135}\times\left(192.39-\frac{109}{135}\times192.39-\frac{26}{135}\times1\,153.94\right)^2+\frac{26}{135}\times\left(1\,153.94-\frac{109}{135}\times192.39-\frac{26}{135}\times1\,153.94\right)^2=143\,772.6.$$

进一步计算可得

$$K=\frac{v}{a}=0.363\,6,$$

$$Z=\frac{T}{T+K}=\frac{135}{135+0.363\,6}=0.997\,313\,9.$$

经计算可得一维情况中的布尔曼因子约为 0.99，这表示样本数据中个别数据点对最终类别的影响极大. 此时，如果结合了具有强烈偏见的先验参数信息，可能导致所有样本数据最终都被分类到同一类别中.

布尔曼信度因子 Z 在统计决策过程中用于衡量个体数据与整体数据的相对重要性. 一个接近 1 的 Z 值表明个体数据对总体分类结果的影响非常大. 当这种高信度因子与先验信息结合时，如果先验信息本身具有将数据归类到某一特定类别的倾向，这种倾向会被放大，从而导致所有数据都倾向于被分到这一类.

高信度因子可能导致分类模型过度适应(过拟合)到某些特定的样本数据特征上，尤其是当这些数据与强先验信息一致时. 这种情况下的分类结果虽然可能在训练集上表现良好，但在新的或未知的数据上可能无法保持同样的表现，因为模型缺乏泛化能力.

当所有样本数据都被分到同一类别中时，可能忽视了数据内在的多样性和复杂性. 这种做法在面对具有不同特征或属于真实多类的新数据时，可能导致错误的分类决策，增加决策的风险.

在实际应用中，需要对 Z 值进行仔细的调整和管理，以平衡先验信息和样本数据的影响. 适当降低 Z 值，使得决策过程更多地考虑到数据的整体分布特性，有助于提高模型对新情况的适应能力和决策的可靠性.

4 总结与展望

本文基于混合 Gaussian 分布和 EM 算法提出了一种高效的数据参数估计和动态分类方法. 该方法不仅深入剖析了各品类的详细特征，更通过精准的数据分析，为客户提供全面、详尽的决策支持. 特别值得一提的是，它能够精确计算出每个品类在各个类别中的概率分布，为市场策略的定制、产品组合的优化提供坚实的数据基础.

在品类分类的过程中，应秉持科学、客观的统计学原则，确保分类结果的准确性和公正性. 同时，考虑到商业环境的复杂多变，允许根据实际情况灵活调整品类分类更有利于适应市场的快速变化. 这种灵活性使得本文方法更具实用性，能够更有效地满足客户的多样化需求.

从理论层面，详细推导了参数更新的数学公式，使得算法更加精准地适应各类问题的求解. 通过模拟实验和实证分析，验证了该方法在处理混合分布数据时的卓越性能，并展示了其在实际应用中的巨大潜力. 这些研究不仅有助于理解市场数据，还为销售策略的制定提供了有力的数据支撑.

此外，本研究不仅为混合分布模型和数据分类领域带来了新的发展思路，也为相关行业的数字化转型和智能化升级提供了有力的支持. 特别是，通过引入品类概率分布的概念，为数据分析注入了更多的灵活性和实用性，使其成为了企业和组织在激烈市场竞争中脱颖而出的关键工具.

尽管当前研究取得了初步成果，但未来仍有多个方向值得进一步探索. 首先，可以考虑采用混合 Erlang 分布等具有优良性质的模型来拟合数据，以简化模型复杂度同时保持拟合效果. 其次，针对 EM 算法对初始值敏感的问题，研究如何设定合适的初始值以加快算法收敛速度和提升优化效果. 再者，将信度因子推广至高维情形，以更准确地衡量高维数据的风险特征. 同时，探讨在初始概率模型中考虑 θ 连续分布的情况，以扩展模型的应用范围.

此外，将本研究方法与其他领域问题结合，如解决团簇质谱数据分析的相关问题，求解组成成分个

数、具体组成成分、各个成分在团簇中所占比例,如最终的分子式可能为:

$$x_1La^{3+}+x_2NiC_4O_4NH_5+x_3CH_3O^-+x_4O^{2-}+x_5OH^-+x_6H_2O+x_7H^+,$$

在限制条件之下进行求解:

$$\begin{cases}1\leqslant x_1\leqslant 5; & 0\leqslant x_2\leqslant 8;\\ 0\leqslant x_3\leqslant 10; & 0\leqslant x_4\leqslant 10;\\ 0\leqslant x_5\leqslant 10; & 0\leqslant x_6\leqslant 3;\\ 0\leqslant x_7\leqslant 10; & x_i\in Z(i=1,\cdots,7).\end{cases}$$

未来研究应在更广泛的行业和应用背景下测试这些模型的适用性和扩展性,以优化模型结构和参数估计方法,更好地服务于商业分析和决策制定.

参考文献:

[1] 张娣. 变分贝叶斯在混合厄朗模型的应用[D]. 厦门:厦门大学,2021.

[2] THOMAS H, HETTMANSPERGER T P. Modelling change in cognitive understanding with finite mixtures [J]. Journal of the Royal Statistical Society: Series C (Applied Statistics),2001,50(4):435-448.

[3] NEMEC J,NEMEC A F L. Mixture models for studying stellar populations. i. Univariate mixture models, parameter estimation, and the number of discrete population components[J]. Publications of the Astronomical Society of the Pacific,1991,103(659):95.

[4] PAGEL M,MEADE A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data[J]. Systematic Biology,2004,53(4):571-581.

[5] VARDI Y, SHEPP L A, KAUFMAN L. A statistical model for positron emission tomography[J]. Journal of the American Statistical Association, 1985, 80(389): 8-20.

[6] RASMUSSEN C. The infinite Gaussian mixture model[J]. Advances in Neural Information Processing Systems, 1999,12:1-10.

[7] DEMPSTER A P,LAIRD N M,RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological),1977,39(1):1-22.

[8] REDNER R A,WALKER H F. Mixture densities,maximum likelihood and the EM algorithm[J]. SIAM Review, 1984,26(2):195-239.

[9] 杜让. 基于 EM 算法的时间序列分位数回归模型的统计推断[D]. 长春:长春工业大学,2023.

[10] MOON T K. The expectation-maximization algorithm[J]. IEEE Signal Processing Magazine,1996,13(6):47-60.

[11] BILMES J A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models[J]. International Computer Science Institute,1998,4(510):126.

[12] MACQUEEN J. Some methods for classification and analysis of multivariate observations [J]. Berkeley Symposium on Mathematical Statistics and Probability, 1967,1(14):281-297.

[13] ESTER M,KRIEGEL H P,SANDER J,et al. A density-based algorithm for discovering clusters in large spatial databases with noise[J]. Kdd,1996,96(34):226-231.

[14] FRALEY C,RAFTERY A E. Model-based clustering, discriminant analysis,and density estimation[J]. Journal of the American statistical Association,2002,97(458):611-631.

(责任编辑:汪　军)