 DOI: https://doi.org/10.12349/tie.v3i8.11545

A Parameter Determination Method for Financial Complaint Text Clustering Based on Adjusted Hierarchical Homogeneity

Tianli Tao Jingang Shi

School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang, Liaoning, 110168, China

Abstract

Financial complaint text clustering is sensitive to parameter settings. Existing internal indices rarely reflect business semantics and may overrate fragmented partitions as cluster count increases. This paper proposes Adjusted Hierarchical Homogeneity (AHH), which uses two-level business labels as weak supervision, retains the asymmetric direction of V-measure homogeneity, removes the completeness component that discourages legitimate refinement, corrects incidental agreement with expected mutual information, and harmonically aggregates coarse- and fine-level scores. Among 32 CFPB candidates, AHH selects the partition ranked first by the held-out NMI-ARI joint index and outperforms five baselines. It remains competitive on eight additional public datasets, while OnCoCo identifies the applicability boundary caused by a degenerate coarse label. AHH therefore turns experience-based geometric trial and error into reproducible parameter selection driven by business semantics.

Keywords

text clustering; clustering parameter determination; weak supervision; hierarchical homogeneity; fine-grained issue discovery

基于随机校正层级同质性的金融投诉文本聚类参数确定方法

陶天立 师金钢

沈阳建筑大学计算机科学与工程学院，中国 · 辽宁 沈阳 110168

摘　要

金融投诉文本聚类对参数敏感，现有内部指标又难以反映业务语义，并可能因簇数增加而高估过碎划分。为此，提出随机校正层级同质性参数确定方法AHH：以文本自带的两级业务标签为弱监督信号，保留V-measure同质性的非对称方向，舍弃限制细分的完整性分量，引入期望互信息校正偶然一致性，并调和粗、细两层得分。CFPB的32组候选实验中，AHH选出的划分在留出NMI-ARI联合指标上排名第一，优于五种基线；在另外八个公开数据集上总体保持前列，OnCoCo揭示了粗层标签退化时的适用边界。结果表明，AHH可将依赖经验的几何试参转化为业务语义驱动、过程可复现的自动选参。

关键词

聚类评价指标；聚类参数确定；弱监督；细粒度问题发现；双标签

1 引言

金融投诉文本数量大、篇幅短且业务术语密集，是识别服务短板和合规缺陷的重要来源。粗粒度分类难以揭示具体问题，逐条复核成本又高；聚类虽可在缺少细粒度标注时生成候选问题簇，却对降维和密度参数敏感，不同参数常产生粒度悬殊的结果。

短文本与主题聚类研究为投诉文本无监督划分提供了基础[1-2]，密度聚类也常用于生成候选簇[3]。但轮廓系数、Calinski-Harabasz 和 Davies-Bouldin 等内部指标只反映向量空间的紧致性与分离度，难以对应业务边界；V-measure 虽能衡量聚类与标签的一致性，其完整性分量会限制合理细分，未校正的同质性还可能随簇数增加而虚高。

本文将已有的两级业务标签作为弱监督信号，提出 AHH。该方法保留 V-measure 同质性的非对称方向，舍弃完整性，利用期望互信息扣除偶然吻合，并调和粗、细两层得分。CFPB 主实验比较 AHH 与五种基线，消融实验考察随机校正和层级设计，跨数据集实验验证其稳定性与适用边界。

2 随机校正层级同质性方法

2.1 问题定义

给定文本句向量集合和候选参数空间 Θ，参数 θ 对应聚类划分 $C(\theta)$，噪声记为 −1。目标是选出 θ^*，使 $C(\theta^*)$ 既保持簇内业务语义纯度，又避免同类问题被切散为过多小簇。两项目标分别偏好细、粗划分，需要在其间确定可复核的粒度。本文以文本自带的粗层标签 Y1 和细层标签 Y2 作

【作者简介】陶天立（1994–），女，中国江苏南京人，在读硕士，审计师，从事文本挖掘与聚类分析研究。

为弱监督信号。

2.2 经典 V–measure 同质性及其缺陷

V-measure 由同质性与完整性的调和平均构成。同质性衡量每个簇是否主要来自同一标签，其定义如式 (1) 所示。

$$h(Y,C) = 1 - \frac{H(Y|C)}{H(Y)} = \frac{I(Y;C)}{H(Y)} \tag{1}$$

式中 H(Y) 为标签熵，H(Y|C) 为条件熵，I(Y;C) 为互信息。同质性以标签为被解释对象、聚类为解释变量；若每个簇只含同一标签，H(Y|C)=0，得分为 1。这一非对称方向符合“给定一个簇判断其业务归属”的复核需求。

经典 V-measure 存在两点不足：完整性要求同一标签尽量集中于同一簇，会将比输入标签更细的合理划分误判为碎裂；互信息未作随机校正，大量小簇即使与标签无真实关系，也会因偶然吻合取得正值，从而系统性高估过碎方案。CFPB 消融中，去掉随机校正的双层同质性最终选择 1173 簇，留出联合排名降至第 6。

2.3 随机校正同质性

为校正第二项偏差，在固定标签与簇边际的条件下，按超几何模型计算调整互信息所用的期望互信息 E[I(Y;C)]，见式 (2)。

$$E[I(Y;C)] = \Sigma_i \Sigma_j \Sigma_{nij} \frac{n_{ij}}{N} \cdot \ln\left(\frac{N \cdot n_{ij}}{a_i \cdot b_j}\right) \cdot P(n_{ij}) \tag{2}$$

其中，i=1,⋯,p 和 j=1,⋯,m 分别为标签类别与聚类簇索引；$a_i=\sum_j n_{ij}$、$b_j=\sum_i n_{ij}$ 为边际计数，$n_{ij}=|Y_i \cap C_j|$，且 $\sum_i \sum_j n_{ij}=N$。$P(n_{ij})$ 为固定行列边际下该单元服从超几何分布的概率。E[I] 可解析计算，无需随机模拟，并随候选划分的簇数和规模结构变化。据此定义随机校正同质性，见式 (3)。

$$ha(Y,C) = \max\left\{0, \frac{I(Y;C) - E[I(Y;C)]}{H(Y) - E[I(Y;C)]}\right\} \tag{3}$$

式 (3) 以超过随机基线的互信息为分子，以从随机水平到完美保持标签的剩余空间为分母，使随机水平和完美同质分别对应 0 和 1。令 b=E[I(Y;C)]/H(Y)，可改写为式 (4)。

$$ha(Y,C) = \max\left\{0, \frac{h - b}{1 - b}\right\} \tag{4}$$

式 (4) 表明，校正先从同质性中减去当前簇数与簇规模造成的随机基线 b，再按剩余空间归一化。只有超过 b 的部分才表示候选划分真正保持的业务边界；max 将低于随机水平的值截为 0，避免负值干扰排序。

2.4 双层业务标签与 AHH 定义

针对第一点缺陷，本文弃用完整性分量，改为在两个业务粒度上分别计算同质性。粗层取 Y1，约束簇不跨越明显的业务大类；细层取 Y2，进一步保证簇内样本在具体业务子类上尽量一致。两层分量如式 (5) 所示。

$$ha1 = ha(Y1, C), \quad ha2 = ha(Y2, C) \tag{5}$$

AHH 得分取两层随机校正同质性的调和平均，如式 (6) 所示。

$$AHH(\theta) = \frac{2 \cdot ha1 \cdot ha2}{ha1 + ha2} \tag{6}$$

调和平均不设数据集相关权重，并要求两层同时较高；任一层接近 0，AHH 都会显著下降。粗层同质性约束簇不跨越业务大类，细层同质性保证具体子类纯度，随机校正则抑制碎分虚高。AHH 与 V-measure、AMI 同源但含义不同：它舍弃 V-measure 的完整性，只回答聚类在多大程度上保持两级业务标签边界；其归一化以标签熵 H(Y) 为基准，而非 AMI 使用的对称熵。

2.5 选参算法

遍历 Θ 中的每个 θ，执行降维和聚类得到 C(θ)，分别构造其与 Y1、Y2 的列联表，计算两层校正同质性及 AHH，最后按式 (7) 取最大值。AHH 仅依赖列联表、熵和期望互信息，复杂度约为 O(N+m(p+q))，其中 N 为有效样本数，m 为簇数，p、q 为两层标签类别数；它不需重采样，可直接嵌入批量选参流程。

$$\theta^* = \arg\max_{\theta \in \Theta} AHH(\theta) \tag{7}$$

3 实验与分析

3.1 数据集与实验设置

主实验采用 CFPB 2024 年公开金融投诉正文，分层平衡抽样 12750 条，产品、子产品及“问题—子问题”路径分别形成 9、39、227 类；清洗后词数 P50/P95 为 156/642，采用 bge-m3 编码。Cat1 产品和 Cat2 子产品作为两级弱监督信号，Cat3 仅用于留出评价。跨数据集另取 Amazon、DBpedia、Shopify、Essential-Web、BestBuy、WOS-JTF、HUPD 和 OnCoCo 八个公开数据集；除 OnCoCo 外均有有效的三级语义层级。

32 组候选由两套网格生成：HDBSCAN 的 n_neighbors 取 5、10、15、30，min_cluster_size 取 2、3、5、8、12、20、40、80；OPTICS 沿用 n_neighbors，min_samples 取 2、5，噪声率取 0.1、0.2、0.3、0.4，并按可达距离分位数定 eps。UMAP 统一降至 50 维（min_dist=0），用余弦距离，在非噪声样本上评价。留出指标为 NMI、AMI、ARI，联合指标 J 按式 (8) 计算：M ∈ {NMI,ARI} 在候选池内作 min-max 归一化，分母为 0 记 0。AHH 只读取 Cat1/2，定 θ* 后用 Cat3 评价。

$$\hat{M}\theta = \frac{M\theta - \min_{\theta' \in \Theta} M\theta'}{\max_{\theta' \in \Theta} M\theta' - \min_{\theta' \in \Theta} M\theta'}, \quad J\theta = \sqrt{NM\hat{I}\theta \cdot AR\hat{I}\theta} \tag{8}$$

3.2 CFPB 主实验：选参方法对比

表 1 给出 CFPB 的六种 HDBSCAN 选参结果。AHH 选中 nn=30、mcs=3，得到 525 个非噪声簇，覆盖率 0.5668；留出 NMI、AMI、ARI 分别为 0.5835、0.3120 和 0.1196，J=0.9685，在 32 个候选中排名第一。轮廓系数与 Calinski-Harabasz 指数均选 55 簇，联合第 8；Davies-Bouldin 指数偏向 881 簇，列第 19；相对 DBCV 和 Cat1 V-measure 分别列第 29、第 28。

表 1 CFPB 数据上六种方法的 HDBSCAN 选参结果

选参准则	nn/mcs	K	覆盖率	NMI	AMI	ARI	联合指标值	联合排名
轮廓系数	30/40	55	0.5605	0.5009	0.4054	0.1129	0.8719	8
Calinski-Harabasz	30/40	55	0.5605	0.5009	0.4054	0.1129	0.8719	8
Davies-Bouldin	5/3	881	0.7140	0.5805	0.2358	0.0847	0.8127	19
相对 DBCV	10/40	46	0.7490	0.4478	0.3810	0.0637	0.6193	29
Cat1 V-measure	30/80	25	0.6689	0.4604	0.4150	0.0629	0.6239	28
AHH（本文）	30/3	525	0.5668	0.5835	0.3120	0.1196	0.9685	1

几何指标选出的粒度跨度为 55—881 簇，表明向量空间的紧致与分离难以稳定对应业务边界；AHH 舍弃完整性并校正随机增益后，在同一候选池中取得更好的外部折中。

3.3 消融实验

在相同候选网格下，分别比较去随机校正的 UAHH、仅用粗层的 CoarseAH、仅用细层的 PathAH、条件增量同质性 CAHH 和 AHH。UAHH 与 CAHH 均选择 K=1173 的过碎划分，联合第 6；PathAH 选择 368 簇、列第 2；CoarseAH 与 AHH 均选择 525 簇、列第 1。结果表明，随机校正是抑制簇数膨胀的关键，双层调和则提供不依赖单一层级的稳健约束。

3.4 跨数据集泛化验证

在 HDBSCAN 与 OPTICS 两套密度聚类的候选参数上分别用 AHH 选参，再用未参与选参的 Cat3 计算留出指标，结果见表 2。除边界案例 OnCoCo 外，七个具有有效粗细层语义的数据集上，AHH 总体偏强。Amazon 在两套候选上都同时取得留出 NMI 与联合指标第一；WOS-JTF 的 OPTICS 候选四项指标全部第一；Essential-Web、BestBuy 的联合排名进入前四；DBpedia 与 Shopify 落在前八；CFPB 在 HDBSCAN 上联合第一、OPTICS 上联合第二。HUPD 的专利文本极为细碎，留出 Cat3 多达 5947 类、中位每类仅一条，AHH 选中划分的留出 AMI 进入前三，说明其语义纯度高，但裸 NMI 偏低。相比之下 OnCoCo 明显失败，其留出联合排名在两套候选上分别为第 21 与第 32。

表 2 跨数据集 AHH 选参的留出联合排名（32 候选内）

数据集	HDBSCAN 选中 K	HDBSCAN 联合排名	OPTICS 选中 K	OPTICS 联合排名
Amazon	940	1	897	1
CFPB	525	1	1142	2
DBpedia	362	8	342	6
Shopify	339	5	437	3
Essential-Web	172	2	412	10
BestBuy	861	3	838	4
WOS-JTF	279	7	104	1
HUPD	413	9	790	5
OnCoCo	343	21	465	32

3.5 讨论：NMI 簇数偏置与适用边界

细粒度选参必须正视评价指标自身的簇数偏置。NMI 以互信息为分子，而互信息随簇数增多单调不减，划分越细，NMI 越可能虚高。若直接以 NMI 最大化选参，往往偏向过碎划分。ARI 与 AMI 经随机一致性校正，对过碎划分更稳健。本文因此以 NMI 与 ARI 的联合指标而非单一 NMI 作为留出主口径。AHH 本身的随机校正正对应 NMI 偏置的成因，使其在候选排序上比裸 NMI 更能抵抗碎分。

AHH 的有效性有明确前提。其一，需要至少两级、与问题内容相关且不过度退化的元数据。OnCoCo 的 Cat1 仅 2 类，且按说话人角色而非对话内容划分，粗层同质性饱和、随机校正的区分作用减弱，AHH 便退化为由细层单独驱动的过度碎分，这正是它失败的原因。其二，AHH 更适用于 HDBSCAN、OPTICS 这类密度聚类候选，因为它们的簇对应特征空间中的局部密集区域，较易与业务层级对齐；对 Leiden 这类基于图模块度的无噪声社区发现，其适用性仍需进一步验证。其三，AHH 是选参方法而非搜索框架，作用是在给定候选集合中排序，仍需候选聚类器在过细划分时表现出可观察的噪声率或簇数约束。

4 结论

针对金融投诉文本细粒度问题发现中的参数敏感问题，本文提出随机校正层级同质性参数确定方法 AHH。它保留 V-measure 同质性的非对称解释方向，弃用会锁死粒度的完整性分量，引入调整互信息式随机校正，并在两级业务标签上作无自由权重的调和聚合。CFPB 主实验中，AHH 选中的划分在留出联合指标上排名第一，优于轮廓系数、Calinski-Harabasz、Davies-Bouldin、相对 DBCV 与单层 V-measure 等基线方法；消融实验表明随机校正能抑制过碎选择；跨数据集实验在 CFPB 之外的八个公开数据集上进一步检验，其中七个具备有效粗细层语义的数据集总体排名靠前，OnCoCo 则界定了粗层退化时的失效边界。AHH 把依赖经验的几何试参转化为业务语义驱动、过程可复现的自动选参，其前提是数据具有至少两级有效元数据。后续将在更多带层级元数据的场景中验证，并研究适用于无噪声候选的粒度约束机制。

参考文献

[1] 贺文灏, 吴春江, 周世杰, 等. 融合无监督SimCSE的短文本聚类研究[J]. 计算机科学, 2023, 50(11): 71-76.

[2] 徐菲菲, 陈赛红. 中文文本主题聚类算法研究综述[J]. 上海电力大学学报, 2021, 37(6): 613-619.

[3] 苏桂娜. 基于标准化自适应参数的密度聚类算法研究[J]. 电子技术与软件工程, 2023(3): 247-251.