



基于 A-D 模型的 K-means 算法 在通话异常客户挖掘中的应用

周坚¹, 石永革¹, 何美斌²

(1. 南昌大学信息工程学院, 江西 南昌 330029;

2. 中国电信股份有限公司江西分公司, 江西 南昌 330029)

摘要: 为能够利用海量的语音通信记录, 高质量地对各种语音通信行为异常的客户 (电信诈骗客户、广告客户等) 进行聚类分析, 设计构建了语音通信异常客户的行为特征模型, 并基于该模型提出了一种语音通信行为异常客户的聚类分析算法。首先, 通过分析客户的通话记录得出客户通话次数、接通率等通话行为特征, 然后将 AHP 模型与 DEMATEL 方法融合, 构建语音通信行为异常客户的行为特征模型 (AHP-DEMATEL 模型); 其次, 基于该模型提出了一种改进的 K-means 算法, 实现根据语音通信记录对异常客户进行聚类分析。最后, 使用真实数据进行了验证分析。结果表明, 相较于其他类似算法, 本文算法在多类型异常客户综合聚类分析和单类型异常客户聚类分析时, 其性能都得到了较大幅度的提高。

关键词: 语音通信; 异常客户挖掘; 行为特征分析; AHP; DEMATEL

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018021

Application of K-means algorithm based on A-D model in calling abnormal customer mining

ZHOU Jian¹, SHI Yongge¹, HE Meibin²

1. Information Engineering College, Nanchang University, Nanchang 330029, China

2. Jiangxi Branch of China Telecom Co., Ltd., Nanchang 330029, China

Abstract: In order to make use of massive voice communication records and cluster high-quality clients (telecom fraud clients, advertisers) with various kinds of voice communication abnormalities, a behavioral feature model of abnormal voice communication customers was designed and constructed. Based on the model, a clustering analysis algorithm for customers with abnormal voice communication behavior was proposed. First of all, by analyzing the call records of customers, the characteristics of call behaviors was got, such as the number of calls, call rates, and so on. Then AHP-DEMATEL model was constructed by blending the AHP model and DEMATEL method. Secondly, based on the model, an improved K-means algorithm was proposed to cluster the abnormal clients according to the voice communication records. Finally, the real data was used to verify the analysis. The results show that compared with other similar algorithms, the proposed algorithm improves the performance of multi-type abnormal customer comprehensive clustering analysis and single-type abnormal customer clustering analysis greatly.

Key words: voice communication, abnormal customer mining, behavior characteristics analysis, analytic hierarchy process, decision making trial and evaluation laboratory



1 引言

随着通信技术的迅速发展,移动和固定电话已经成为人们日常生活中不可缺少的语音通信工具。然而,其在给人们的工作、生活带来极大便利的同时,也会给电信客户和运营商带来负面影响,例如电信诈骗、广告骚扰、欠费等。目前,针对语音通信行为异常客户(简称异常客户),相关部门已经采取了若干措施,如针对电信诈骗实行手机实名制、针对欠费客户的分析催缴等,但其一般都属于事后被动处置。为此,迫切需要基于海量的语音通信数据,自动分析它们之间的关联关系,聚类分析出其中隐藏的异常客户信息,对异常客户及相关事件实现预判预警、实时提示、主动响应、快速处置。

迄今,国内针对电信客户语音通信行为特征进行分析建模并进一步实现客户聚类分析的文献不多,典型如:参考文献[1]采用模糊集的语义关联度算法,可以找出具有强关联度的客户;参考文献[2]应用复杂网络理论和模糊决策树算法,基于通话记录开展语音通信行为异常客户的识别研究;参考文献[3]研究如何使用决策树分析恶意欠费客户,应用聚类分群发现电信欺诈客户。上述文献的主要不足在于,都没有进行深入的异常行为特征建模和客户聚类挖掘算法的研究。参考文献[4]基于客户通话数据进行聚类分析,获取具有良好特征的客户集合,然后应用 *K-means* 算法挖掘特定的客户,以便针对不同客户群提供差异化服务,其主要不足在于:没有对挖掘算法进行优化,使得其挖掘效率和质量不够理想,并且仅考虑了移动通信客户。综上所述,参考文献[1-4]对于本文具有一定的参考价值,但它们的研究或者仅局限于行为特征分析,或者局限于客户挖掘,并没有结合行为特征分析建模与客户挖掘,开展完整深入的研究及算法优化工作,使得其在准确性、完整性、普适性等综合性能上存在明显不足。

参考文献[5-7]对本文具有较高的参考价值,都对聚类算法进行了改进并应用于电信客户细分中。

其中,参考文献[5]通过改进的 *K-means* 算法对电信客户进行了聚类分析,但其主要研究目标是降低聚类分析的时间复杂度。参考文献[6]使用自适应差分演化算法改进 *K-means* 算法,并将其应用在电信客户细分中,但未给出 3 个关键参数的经验值和选取方法。参考文献[7]通过不断优化新的聚类中心改进 *K-means* 算法,从而提高其在电信 CRM (customer relationship management, 客户关系管理) 客户分类中的准确性。

为实现对语音通信行为异常客户的高质量聚类分析,本文针对固定和移动电话客户,以 AHP 模型为基础,融合 DEMATEL 法构建异常客户行为特征模型,即 AHP-DEMATEL 模型 (analytic hierarchy process decision making trial and evaluation laboratory, 以下简称 A-D 模型),并提出了一种基于 A-D 模型的 *K-means* 算法,从海量的语音通信数据中聚类分析出异常客户。本文主要工作是:基于语音通信记录,详细分析客户的通话行为特征;基于 A-D 模型构建语音通信异常客户行为特征模型,以获取各项行为特征的权值;对 *K-means* 算法进行分析和优化,提出基于 A-D 模型的异常电信客户挖掘算法;对上述研究成果进行测试、比较、分析,验证其有效性。

2 相关工作

2.1 层次分析法 AHP

AHP 一种将定性和定量分析融合的多要素决策方法,其主要思想为:针对复杂决策问题,综合其主要影响元素、问题本质、内在关系等,通过深入分析来构造层次结构模型,从而基于很少的定量信息就可将决策者的思维过程数学化。由此,即可为求解拥有多目标、多准则、无结构等特性的复杂决策性问题,提供一种较为简单的决策方法^[8]。

AHP 常用于主观赋权,其具体步骤如下。

步骤 1 构造层次结构模型。将复杂的多目标、多准则的决策问题,按照它们之间的相互关系分解成多个层次,自顶向下每一层对应一组分目标、准则直至相关指标(统称元素),形成递阶层次结构模型。由此,即可通过逐层向上分析,最终确定各低层元素对于各级高层元素的重要性权值。

步骤 2 构建判断矩阵。为尽可能减少性质不同元素相互比较的困难,提高准确度,对于每一个目标,不是将影响该目标的所有元素放在一起比较,而是两两相互比较后,给出描述相对重要程度的比较值。由此,对于各个目标,生成一系列的两两判断矩阵 $A = (a_{ij})_{nm}$, 其中 a_{ij} 表示针对同一目标,元素 i 比 j 的相对重要值。例如,若对于目标 Y , $a_{ij}=2$ 表示元素 i 比元素 j 对 Y 的影响更大(即权重更高),且对应必有 $a_{ji}=1/2$ 。上述重要值应由多位专家进行评价并取其加权均值。

步骤 3 验证判断矩阵的一致性。由于判断矩阵的值主要取决于专家意见,难免出现逻辑不一致的情况(如前例中,若 $a_{ij}=2$, 而 a_{ji} 不是 $1/2$),故必须对判断矩阵进行一致性检验,方法为:首先计算出判断矩阵的最大特征值 $\lambda_{\max}$,利用式(1)计算出一致性检验的标度,然后计算随机一致性比率 CR ,即 $CR = CI/RI$ 。其中, RI 可查表得到,一般认为 $CR < 0.1$ 时具有满意的一致性。

$$CI = (\lambda_{\max} - n) / (n - 1) \quad (1)$$

步骤 4 确定权重。一般采用求和法,按式(2)将判断矩阵的每列向量归一化,再对行求和并归一化。

$$W = a_{ij} / \sum_{i=1}^n a_{ij} \quad (2)$$

2.2 DEMATEL 法

DEMATEL (decision making trial and evaluation laboratory) 即“决策实验和评价实验法”,是一种用于筛选复杂系统主要元素、简化系统结构分析过程的方法论,可以充分利用专家经验和知识处

理比较复杂的系统问题,特别是对关系要素不太确定的系统更为有效^[9]。实施步骤如下。

步骤 1 使用德尔斐法、头脑风暴法或者专家访谈法确定各个元素。

步骤 2 确定各要素间的直接影响度。首先使用专家评分法建立各风险影响元素的直接影响矩阵,一般将元素之间的关系定性地分为 5 个等级,0 分表示无影响关系,1 分表示关系很弱,2 分表示关系较弱,3 分表示关系较强,4 分表示影响关系很强。然后,使用上述评分结果建立直接影响矩阵 $A = (a_{ij})_{nm}$, 其中 a_{ij} 即上述评分值。

步骤 3 将直接影响矩阵规范化。利用 $G = A / \max_{1 \leq i < n} \sum_{j=1}^n a_{ij}$ 计算出规范化直接影响矩阵 G 。

步骤 4 构建综合影响矩阵。利用 $T = G(I - G)^{-1}$ 计算出规范化直接影响矩阵,其中 I 为单位矩阵,为 $(I - G)$ 的逆矩阵。

步骤 5 获取各要素的影响度和被影响度。利用式(3)和式(4)可以分别计算出风险元素之间的相互影响度和被影响度。其中, f_i 是综合影响矩阵 T 中行元素之和,表示风险元素 i 给 j 带来的直接或间接影响度, e_i 是综合影响矩阵 T 中列元素之和,各行对应风险元素受到其他各元素的影响值称为被影响度。

$$f_i = \sum_{j=1}^n t_{ij}, i=1,2,\dots,n \quad (3)$$

$$e_i = \sum_{j=1}^n t_{ji}, i=1,2,\dots,n \quad (4)$$

步骤 6 计算各要素的中心度。使用式(5)即可计算出各要素的中心度。

$$m_i = f_i + e_i, i=1,2,\dots,n \quad (5)$$

2.3 K-means 算法

K-means 算法的主要优点:聚类速度快、效率高、适用于处理大数据集,是一种常用的无监督学习方法。算法主要原理:随机选出 K 个点用作初始聚类中心,然后计算各个样本与初始聚类



中心之间的距离,并且把样本划分到距离最近的聚类中心所在类;对划分完全后的新类进行计算得到新聚类中心,如果前后两次的初始聚类中心未发生变化,可以确定样本调整结束,此时相应的误差平方和函数已达到最小,聚类准则函数已经收敛;否则,若初始聚类中心发生变化,则需要继续调整所有样本来确定新的聚类中心,并进入下一次的迭代过程^[10]。

K-means 算法的主要不足:对初始中心敏感;无法处理非球形类;对孤立点数据和噪声数据较敏感;对初始值 k 的选取依赖性较大;容易陷入局部最优解;同等看待每一个特征,难以产生高质量的聚类结果。

3 语音通信异常行为分析模型

3.1 语音通信行为特征

电信运营商提供的语音通信原始通话记录组成客户原始数据集,其每条记录包括 18 个字段,从中析取开始时间、主叫号码、被叫号码、呼叫结果、通话时长等作为原始特征字段并按客户归并后,即可生成客户通话数据集。

基于语音通信行为异常客户行为特征的分析,需要对客户通话数据集中的数据基于客户进行以下处理后,生成客户通话特征集,其中每一客户的语音通信行为特征参数包括以下几种。

- 平均呼出时间:客户每日呼出时间点的平均值 c_1 ;
- 平均通话时长:客户接通记录中通话时长的平均值 c_2 ;
- 呼出次数:客户主叫次数 c_3 ;
- 呼出号码数:客户主叫时对端号码不重复的个数 c_4 ;
- 呼入次数:客户被叫记录数 c_5 ;
- 呼入号码数:客户被叫时对端号码不重复的个数 c_6 ;
- 接通率:客户主叫接通数与所有主叫记录

的比值 c_7 。

以广告客户为例,其广告推销行为一般集中在白天,平均呼出时间一般分布在 8:00—18:00 时间段;广告必须面向大量客户,其呼出次数及呼出号码数量较多;广告推销号码的呼入次数和呼入号码数量极少;由于被叫客户的经验、智能手机提示以及陌生主叫号码等原因,其接通率相对较低。

3.2 行为特征分析模型构建

如前所述,K-means 是一种非常适合挖掘语音行为异常客户的算法,但由于其同等看待所有特征,会影响聚类分析的质量。实际上,不同的语音通信行为特征对于异常客户挖掘过程的贡献,往往存在较大的差别,如对于广告客户的行为特征分析,由于平均通话时长与正常客户的差别很小,其影响应该可以忽略。为此,本文引入行为特征模型的概念,获取不同行为特征的权重,以提高异常客户挖掘的效率与质量。参考文献[11]围绕系统评价中的权重计算问题,提出了 AHP-DEMATEL 方法,通过计算初始权重和影响力,并相互融合来获得综合权重。实验表明,该方法可以较为准确地计算出异常客户通话行为特征的权重。

3.2.1 基于 AHP 生成行为特征的主观权重

判断异常客户需要基于各行为特征实现,其初始权重则需由领域专家生成,属于主观权重,而 AHP 是目前广泛使用的一种主观赋权法。

(1) 构建异常客户层次分析模型

本文讨论的异常客户包括诈骗客户、广告客户、欠费客户,分别用 z 、 g 、 q 表示,并用 y 表示包括这 3 类客户的异常客户群。由此,可以构建异常客户层次分析模型如下,其中用实线连接的元素表示它们之间具有关联关系。

之后,由多位领域专家根据经验,给出各低层元素对高层元素的影响程度相对值并对其加权平均,得到异常客户综合判断矩阵 Y 、诈骗客户

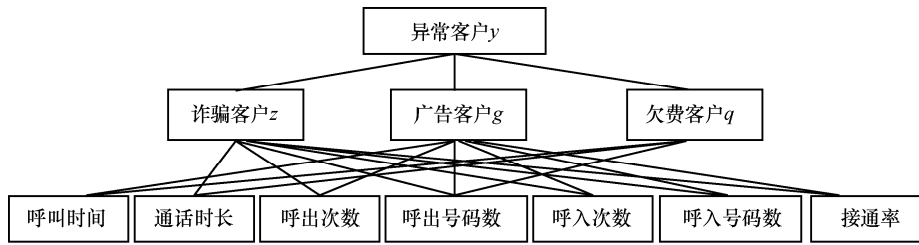


图1 层次分析结构

判断矩阵 Z 、广告客户判断矩阵 G 、欠费客户判断矩阵 Q 。

(2) 一致性检验

通过计算可得矩阵 Y 的 $CR=0.046$ $2<0.10$, Z 的 $CR=0.010$ $3<0.10$, G 的 $CR=0.010$ $3<0.10$, Q 的 $CR=0.028$ $4<0.10$, 所有判断矩阵均可通过一致性检验。

(3) 计算主观权重

通过判断矩阵 Y 得到综合分析异常客户的初始权重向量 W_1 为[0.408 2、0.816 5、0.408 2]。

根据判断矩阵 Z 、 G 、 Q 得到单独分析 3 种异常客户的初始权重矩阵 W_2 见表 1。

表1 3种异常客户分析初始权重矩阵 W_2

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
z	0	0.555	0.555	0.555	0.101	0.101	0.228
g	0.228	0	0.555	0.555	0.101	0.101	0.555
q	0.150	0.916	0	0.370	0	0	0

使用式(2)对 W_2 做归一化处理得到 W_3 , 再使用式(6)即得到异常客户综合分析权重向量 W_4 。

$$W_4 = W_1 \times W_3 \quad (6)$$

对 W_4 归一化处理得到综合权重向量 W_5 。

最后, 将 W_3 和 W_5 组合即得到基于 AHP 的归一化主观权重矩阵 W_A , 见表 2。

表2 基于 AHP 的归一化主观权重矩阵 W_A

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
y	0.088	0.215	0.178	0.233	0.044	0.044	0.195
z	0	0.264	0.264	0.264	0.048	0.048	0.109
g	0.109	0	0.264	0.264	0.048	0.048	0.264
q	0.104	0.637	0	0.257	0	0	0

3.2.2 基于 DEMATEL 确定各特征之间的影响元素

AHP 仅考虑了低层元素对高层元素在垂直方向上的影响, 实际上同层元素之间也存在不同程度的相关性, 从而影响有关权重的取值。如诈骗客户的行为直接关联呼出次数和呼出号码数两项特征, 而它们又会影响到接通率和平均通话时长; 显然, 此时前两项特征的权重应予加强。为此, 本文使用 DEMATEL 方法来进一步提高权重赋值的科学性和客观性。

(1) 根据各特征之间相互影响关系得出直接影响矩阵

采用专家问卷的形式获取相关信息。问卷依次用 0~4 代表各通话行为特征间的关系强度: 无影响、影响稍弱、影响弱、影响强、影响较强。然后, 对收回的有效问卷进行统计整理, 得到直接影响矩阵, 并将该矩阵规范化得到规范影响矩阵 $G(c_i \times c_i)$, 见表 3, 其中, c_i 表示各特征。

表3 规范影响矩阵

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
c_1	0.325	0.793	0.735	0.735	0.583	0.583	0.701
c_2	0.570	0.712	0.992	0.992	0.829	0.829	0.869
c_3	0.584	1.117	1.207	1.430	1.197	1.197	1.193
c_4	0.584	1.117	1.430	1.207	1.197	1.197	1.193
c_5	0.028	0.548	0.861	0.861	0.691	0.913	0.675
c_6	0.028	0.548	0.861	0.861	0.913	0.691	0.675
c_7	0.550	1.015	0.888	1.343	1.343	1.231	0.975

(2) 生成各特征之间的影响度向量

利用规范影响矩阵计算出综合影响矩阵, 并



由其计算出各特征的影响度(R_i)和被影响度(D_i),通过对各特征的影响度和被影响度进行计算,得出该特征的中心度(R_i+D_i),见表4。

表4 中心度

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
R_i	3.193	5.852	7.432	7.432	6.643	6.643	6.283
D_i	4.458	5.796	7.927	7.927	4.84	4.84	7.689
R_i+D_i	7.651	11.378	15.359	15.359	11.483	11.483	13.973

将 DEMATEL 法分析得出的各特征的中心度利用式(2)进行归一化处理,即可得到基于 DEMATEL 的权重向量 $W_D=[0.088,0.131,0.177,0.177,0.132,0.132,0.161]$ 。

3.2.3 生成目标权重

将 W_A 的各行分别与 W_D 进行向量乘运算,再使用式(2)归一化处理后即得到用于异常行为特征分析的目标权重,见表5。

表5 异常行为特征分析的目标权重

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
W_y	0.051	0.185	0.208	0.271	0.038	0.038	0.207
W_z	0	0.218	0.295	0.295	0.040	0.040	0.110
W_g	0.06	0	0.295	0.295	0.040	0.040	0.268
W_q	0.066	0.604	0	0.329	0	0	0

其中, W_y 用于同时综合挖掘异常客户群, W_z 、 W_g 和 W_q 分别用于诈骗客户、广告客户及欠费客户的单类型挖掘。

4 基于 A-D 模型的 K-means 算法

4.1 数据处理

主要工作:从通话记录中析取原始特征字段(开始时间、主叫号码、被叫号码和通话时长等);清洗其中不完整或者不一致的数据;生成客户通话特征集并使用式(7)对所有特征进行归一化处理,得到实验数据集 X 。

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (7)$$

其中, x^* 、 x_i 分别表示归一化前后的值, $x_{\max}$ 、 $x_{\min}$ 分别表示样本数据中的最小、最大值。

4.2 算法描述

为了有效提高语音通信行为异常客户的聚类分析效果,以 K-means 算法为基础,提出基于 A-D 模型的 K-means 算法,主要思路如下。

4.2.1 选择初始聚类中心

聚类算法的基本原理是让类内各数据对象的相似度最小,类间的各数据对象相似度最大,从而提高聚类结果的质量, K-means 算法则由于其波动性而难以保证聚类质量^[15],又由于本文研究所涉及的数据具有海量特征,故本文采用参考文献^[15]提出的方法选择初始聚类中心,方法为:基于异常客户与正常客户相似度较小的原则,不断从最大聚类中找出距离最远的两个数据对象作为新的聚类中心,从而选出 K 个初始聚类中心,消除随机初始聚类中心可能存在的相似度较大问题。

4.2.2 改进欧氏距离公式

K-means 算法是根据数据对象之间的距离进行聚类处理,而数据对象的距离一般采用欧氏距离公式来计算。通过 K-means 常用的欧氏距离公式可见,在计算两个样本的相似度时, K-means 算法将所有特征视为同等重要^[16]。然而,如前所述,不同特征对于样本之间的差异所产生的影响一般是不同的,譬如广告客户关联的 6 项特征中,呼出次数显然更为重要,故在数据样本之间适当放大呼出次数这一特征距离,有利于更精确地分类。由此,本文算法先通过 A-D 模型计算出各特征的最终权重,然后据此适当放大与缩小 K-means 距离公式样本中各特征值之间的距离,使得权重越大的特征对聚类结果的影响越大,从而更加真实地反映实际聚类时各特征的作用。根据上述思路给出的改进距离计算式如下:

$$d(x_i, x_j) = \sqrt{W_1(x_{i1} - x_{j1})^2 + W_2(x_{i2} - x_{j2})^2 + \dots + W_n(x_{in} - x_{jn})^2} \quad (8)$$

其中, d 为所求两个样本之间的距离, x_{in} 表示第 i 个样本的第 n 个特征值, x_{jn} 表示第 j 个样本的第 n 个特征, W_n 表示第 n 个特征的权重。

4.2.3 算法流程

本文算法的形式化描述及相关说明如下。

步骤 1 对数据源进行数据处理, 得到实验数据集, 并从中选出样本集 X 。

步骤 2 基于前述初始聚类中心选择方法选取 k 个初始中心 $R(k)$ 。

步骤 3 生成语音通信行为异常客户挖掘的目标权重: 基于 AHP 法计算不同类型挖掘对象的特征权重矩阵 W_A ; 基于 DEMATEL 法计算各特征间的相互影响权重向量 W_D ; 计算目标权重 W_y 、 W_z 、 W_g 、 W_q 。

步骤 4 选择工作模式 (异常客户综合挖掘, 诈骗客户、广告客户、欠费客户单类型挖掘), 并将相应权值向量代入欧氏距离公式得到 d 。

步骤 5 进行聚类分析, 获取 K 个聚类。

```

for  $k=1, \dots, n$ 
    while 在聚类  $C_k$  中有变化发生
        do 形成聚类;
        for  $k=1, \dots, n$ 
            do  $C_k = \{x \text{ 属于 } X | d(R_k, x) \leq d(R_j, x) \text{ 对所有 } j=1 \dots k, j \neq k\}$ ;
        end;
        计算新的聚类中心;
        for  $k=1, \dots, n$ 
            do  $R_k = C_k$  内点的均值向量;
        end;
    end
end

```

5 实证分析

5.1 实验设置

(1) 实验数据

原始数据来源于某市电信公司的实际语音通信记录, 其中包括 624 199 条固话通话记录、2 611 678 条

手机通话记录, 对应 965 281 位客户。之后, 随机选取 50 000 名客户的数据形成样本集; 析取 100 名正常客户和 100 名异常客户的数据作为测试集 1, 其中平均涵盖了诈骗客户、广告客户和欠费客户; 另析取 100 名正常客户和 100 名广告客户的数据作为测试集 2。

(2) 比对标准

在客户细分分析中, K -means 算法具有较好的效果, 而参考文献[7]通过不断优化聚类中心改进 K -means 算法, 并将其应用于客户细分研究且取得一定效果。因此, 选择 K -means 算法及参考文献[7]提出的改进 K -means 算法进行性能比对。

(3) 评价标准

准确率 (precision) 和召回率 (recall) 是常用性能评价指标, 而鉴于两者往往相互矛盾, 一般采用 F 值 (F -measure) 作为综合评价标准, 相关定义如下:

$$\begin{aligned}
 \text{Precision} &= \frac{TP}{TP+FP} \\
 \text{Recall} &= \frac{TP}{TP+FN} \\
 F\text{-measure} &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Recall} + \text{Precision}} \quad (9)
 \end{aligned}$$

其中, TP 、 FP 、 FN 依次表示正确检测出的异常客户概貌数、错误标识为异常的正常客户数、错误标识为正常的异常客户概貌数。

(4) 参数设置

由于 K 较小时的实验数据不太稳定, 故选 $K=4$ 、5、6、7。

5.2 异常客户综合聚类分析

综合聚类分析指同时对诈骗客户、广告客户、欠费客户 3 类异常客户进行聚类分析。

选用样本集和测试集 1, 得到聚类结果后, 结合电信公司已有数据和人工抽样检测结果可以发现簇内测试集中异常客户多于正常客户时, 该簇



内异常客户数量会明显多于正常客户。因此本实验根据测试集内客户的分布,合并所有异常客户数大于正常客户数的类作为异常客户类,其他为正常客户类。实验结果见表6。

由表6可见,在 $K=5$ 时,本文算法的综合性能F值达到最佳值;通过对3种算法最佳性能比较,可以发现本文算法相对于参考文献[7]算法性能提高4.49%,相对于K-means的提高幅度达到12.05%。此外,本文算法的准确率也明显优于另两种算法。

5.3 广告客户单类型聚类分析

选用样本集和测试集2,异常类合并同上,实验结果见表7。

由表7可见,在 K 值相同时,本文算法的综合性能F值明显优于参考文献[7]算法及K-means,提高幅度分别可达6.52%和7.69%。此外,本文算法的准确率同样明显优于另两种算法。

5.4 小结

综合比较表6和表7的数据,本文算法在广告客户聚类分析时的性能优于多类型异常客户综合聚类分析。本文对诈骗客户和欠费客户进行了类似实验分析,其结果与广告客户聚类分析的结果类似。

因此,单类型异常客户聚类分析的性能优于多类型异常客户综合聚类分析,但综合聚类分析的时间效率则明显高于单类型聚类分析,这种情况符合客观规律。因此,针对不同场景可以选择合适的方案,有利于提高异常客户聚类分析的综合效率。

6 结束语

本文基于语音通信记录,分析了异常客户的行为特征,设计并构建了A-D模型,据此提出了一种改进的K-means算法,实现利用海量数据对语音通信行为异常的客户(电信诈骗客户、广告客户、欠费客户等)进行聚类分析。实证分析结果表明,本文算法用于对语音通信行为异常客户进行聚类分析时的性能效果较好。

参考文献:

- [1] 刘霞. 基于电信通话记录的数据挖掘[J]. 计算机与现代化, 2008(6): 36-38.
LIU X. Application of fuzzy-association rule to telecom data mining[J]. Computer and Modernization, 2008(6): 36-38.
- [2] 王家定. 基于复杂网络理论和通话记录的客户行为异常识别研究[D]. 合肥: 中国科学技术大学, 2013.
WANG J D. Research on customer behavior anomaly recognition based on complex network theory and call logging[D]. Hefei: University of Science and Technology of China, 2013.

表6 异常客户群综合聚类结果分析

K	召回率			准确率			F 值		
	本文算法	K-means	参考文献[7]算法	本文算法	K-means	参考文献[7]算法	本文算法	K-means	参考文献[7]算法
4	0.68	0.56	0.71	0.92	0.91	0.66	0.78	0.69	0.68
5	0.92	0.88	0.83	0.95	0.80	0.83	0.93	0.83	0.83
6	0.91	0.81	0.87	0.91	0.71	0.90	0.91	0.75	0.88
7	0.82	0.82	0.87	0.93	0.77	0.93	0.87	0.79	0.89

表7 广告客户单类型聚类结果分析

K	召回率			准确率			F 值		
	本文算法	K-means	参考文献[7]算法	本文算法	K-means	参考文献[7]算法	本文算法	K-means	参考文献[7]算法
4	0.89	0.82	0.89	0.97	0.84	0.88	0.92	0.82	0.88
5	0.98	0.90	0.92	0.98	0.88	0.88	0.98	0.88	0.89
6	0.99	0.92	0.93	0.98	0.91	0.92	0.98	0.91	0.92
7	0.98	0.91	0.93	0.98	0.91	0.92	0.98	0.91	0.92

- [3] 石瑛. 面向电信市场营销的数据挖掘技术研究[D]. 上海: 华东师范大学, 2004.
SHI Y. Research on data mining technology for telecom marketing[D]. Shanghai: East China Normal University, 2004.
- [4] 赵凯, 蒋朝惠. 基于 K-means 算法的电信客户行为特征聚类分析[J]. 贵州大学学报: 自然科学版, 2015, 32(2): 86-90.
ZHAO K, JIANG C H. Analysis of telecom user behavior features clustering based on K-means algorithm[J]. Journal of Guizhou University(Natural Science), 2015, 32(2): 86-90.
- [5] 江哲雅. 聚类挖掘在电信客户分类中的研究与应用[D]. 上海: 上海交通大学, 2013.
JIANG Z Y. Research and application of clustering mining in telecom customer classification[D]. Shanghai: Shanghai Jiao Tong University, 2013.
- [6] 关云鸿. 改进 K-均值聚类算法在电信客户分类中的应用[J]. 计算机仿真, 2011, 28(8): 138-140.
GUAN Y H. Application of improved K-means algorithm in telecom customer segmentation[J]. Computer Simulation, 2011, 28(8): 138-140.
- [7] ZUO G C, ZHOU R H, LI Z Q, et al. Improved K-means algorithm and its CRM in telecom customer segmentation application[J]. Computer Systems & Applications, 2012, 19(11): 155-159.
- [8] FAN G, ZHONG D, YAN F, et al. A hybrid fuzzy evaluation method for curtain grouting efficiency assessment based on an AHP method extended by D numbers[J]. Expert Systems with Applications, 2016(44): 289-303.
- [9] TZENG G H, CHIANG C H, LI C W. Evaluating intertwined effects in e-learning programs: a novel hybrid MCDM model based on factor analysis and DEMATEL[J]. Expert Systems with Applications, 2007, 32(4): 1028-1044.
- [10] 原福永, 张晓彩, 罗思标. 基于信息熵的精确特征赋权 K-means 聚类算法[J]. 计算机应用, 2011, 31(6): 1675-1677.
YUAN F Y, ZHANG X C, LUO S B. Accurate property weighted K-means clustering algorithm based on information entropy[J]. Journal of Computer Applications, 2011, 31(6): 1675-1677.
- [11] 瞿英, 路亚静, 刘紫玉, 等. 基于 AHP-DEMATEL 法的权重计算方法研究[J]. 数学的实践与认识, 2016, 46(7): 38-46.
QU Y, LU Y J, LIU Z Y, et al. Research on the weight calculation method based on the AHP-DEMATEL method[J]. Mathematics in Practice and Theory, 2016, 46(7): 38-46.
- [12] 丁晓琴, 张德生. 基于 AHP 和 CRITIC 综合赋权的 K-means 算法[J]. 计算机系统应用, 2016, 25(7): 182-186.
DING X Y, ZHANG D S. K-means algorithm based on synthetic weighting of AHP and CRITIC[J]. Computer Systems & Applications, 2016, 25(7): 182-186.
- [13] 王俊凤, 闫文. 基于 AHP-DEMATEL 模型的黑龙江省农民专业合作社内部资金互助的影响因素分析[J]. 金融理论与实践, 2016(7): 86-90.
WANG J F, YAN W. Analysis of the influence of mutual funds between farmer cooperatives in Heilongjiang province based on AHP-DEMATEL model[J]. Financial Theory and Practice, 2016(7): 86-90.
- [14] HUNG S J. Activity-based divergent supply chain planning for competitive advantage in the risky global environment: a DEMATEL-ANP fuzzy goal programming approach[J]. Expert Systems with Applications, 2011, 38(8): 9053-9062.
- [15] 陈光平, 王文鹏, 黄俊. 一种改进初始聚类中心选择的 K-means 算法[J]. 小型微型计算机系统, 2012, 33(6): 170-173.
CHEN G P, WANG W P, HUANG J. Improved initial clustering center selection method for K-means algorithm[J]. Journal of Chinese Computer Systems, 2012, 33(6): 170-173.
- [16] 熊平, 顾霄. 基于特征权重最优化的 K-means 聚类算法[J]. 微电子学与计算机, 2014(4): 40-43.
XIONG P, GU X. K-means clustering algorithm based on feature weight optimization[J]. Microelectronics Computer, 2014(4): 40-43.
- [17] 刘春, 邹海锋, 向勇. 大数据环境下电信数据服务能力开放研究[J]. 电信科学, 2014, 30(3): 156-161.
LIU C, ZOU H F, XIANG Y. Research on telecom data service open ability under the environment of big data[J]. Telecommunications Science, 2014, 30(3): 156-161.
- [18] 韩晶, 张智江, 王健全, 等. 面向统一运营的电信运营商大数据战略[J]. 电信科学, 2014, 30(11): 154-158.
HAN J, ZHANG Z J, WANG J Q, et al. The unified-operation-oriented big data strategy for telecom operators[J]. Telecommunications Science, 2014, 30(11): 154-158.
- [19] 曹旭, 曹瑞彤. 基于大数据分析的网络异常检测方法[J]. 电信科学, 2014, 30(6): 152-156.
CAO X, CAO R T. Network anomaly prediction method based on big data[J]. Telecommunications Science, 2014, 30(6): 152-156.
- [20] 谷红勋, 杨珂. 基于大数据的移动用户行为分析系统与应用案例[J]. 电信科学, 2016, 32(3): 139-146.
GU H X, YANG K. Mobile user behavior analysis system and applications based on big data [J]. Telecommunications Science, 2016, 32(3): 139-146.

[作者简介]



周坚 (1993-), 男, 南昌大学信息工程学院硕士生, 主要研究方向为数据挖掘、计算机网络。

石永革 (1953-), 男, 南昌大学信息工程学院教授, 主要研究方向为计算机网络、信息安全。

何美斌 (1977-), 男, 中国电信股份有限公司江西分公司高级工程师, 主要研究方向为网络信息安全、云计算及大数据分析。