Advances in Applied Mathematics 应用数学进展, 2022, 11(4), 1541-1548
Published Online April 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.114168

Hans汉斯

医保欺诈行为的主动发现

——基于熵权法引入指标权重的聚类分析算法

周彦名

沈阳航空航天大学，辽宁 沈阳

收稿日期：2022年3月8日；录用日期：2022年3月31日；发布日期：2022年4月12日

摘　要

本文针对一些现有的识别方法存在的问题进行改进，应用“基于熵权法引进指标权重的聚类分析算法”进行医保欺诈行为的识别。在完全舍弃主观赋权的情况下，精确地识别出发生欺诈行为的个案。首先对无意义的数据进行降维，并结合医疗保险欺诈的现实案例，综合筛选出五个指标，而后引入信息熵的概念，并基于熵权法确定指标权重；为了避免各个指标的权重给定中存在的主观性，本文通过对信息熵的刻画来体现某一个指标所拥有的信息量期望，最后将得到的权重应用于“改进的欧式距离”，通过对不同指标的“距离”进行赋权，得到一种全新的“距离”用于聚类分析。并按照账单号合并多条拿药记录，以账单号为索引，通过层次聚类分析算法构建聚类树。本文认定：医保欺诈行为是完全地呈孤立点分布的。通过改变聚类数，得到不同聚类数下的孤立点个数，最终结合相关案例，选定聚类数为4，由此求得并给出疑似发生医保欺诈的账单记录43个。

关键词

指标权重，聚类分析，欺诈识别，信息熵，熵权法

Proactive Detection of Health Insurance Fraud

—Clustering Analysis Algorithm Based on Entropy Weighting Method Introducing Indicator Weights

Yanming Zhou

Shenyang Aerospace University, Shenyang Liaoning

Received: Mar. 8th, 2022; accepted: Mar. 31st, 2022; published: Apr. 12th, 2022

文章引用: 周彦名. 医保欺诈行为的主动发现[J]. 应用数学进展, 2022, 11(4): 1541-1548.
DOI: 10.12677/aam.2022.114168

Abstract

This paper addresses the problems of some existing identification methods and applies a "clustering analysis algorithm based on entropic weighting to introduce indicator weights". In order to avoid subjectivity in the weighting of each indicator, this paper uses the information entropy to characterise the expected information content of a particular indicator. In order to avoid the subjectivity in the weighting of each indicator, this paper reflects the information expectation of a certain indicator through the portrayal of information entropy, and finally applies the obtained weights to the "improved Euclidean distance", and obtains a new "distance" by assigning weights to the "distances" of different indicators for clustering analysis. The paper also merges multiple medication taking records by billing number and uses the billing number as the index to construct a clustering tree by a hierarchical cluster analysis algorithm. This paper concludes that: health insurance fraud is completely distributed in isolated points. By varying the number of clusters, the number of isolated points under different clusters was obtained, and finally the number of clusters was selected to be 4 in conjunction with relevant cases, which resulted in 43 billing records suspected to have been fraudulent.

Keywords

Indicator Weights, Cluster Analysis, Fraud Identification, Information Entropy, Entropy Method

1. 引言

医疗保险欺诈，是指公民在参加医疗保险、缴纳医疗保险费、享受国家医疗保险等福利待遇过程中，故意捏造事实、弄虚作假、隐瞒真实情况而导致医疗保险基金损失的行为。孙梦秋[1]指出：骗保人在进行医保欺诈时主要的行为，一是使用他人的医保卡配药，二是在不同的医院或医生处重复配药。为维护国家医疗保险基金，结合现今的大数据时代，如何在海量的数据中寻找到可能发生医保欺诈行为的个案是具有现实意义的。

目前，针对医保欺诈识别的模型，大多采用统计回归、神经网络等辅助(有监督)的学习方法，抑或是非辅助(无监督)学习等。狄萱[2]使用了孤立森林与随机森林的无监督学习方法，刘蓥[3]使用了基于深度学习的有监督学习方法，李金灿、徐珂琳等人[4]总结了应用大数据技术的各类算法，陈富秋、吕亚兰等人[5]证明了依托数据分析对医保异常行为检测的可行性。

诸如机器学习，神经网络等辅助学习(有监督)方法在机器进行学习步骤前，需要人工筛选出一定量的欺诈数据作为训练集，因此存在较强的主观性、工作量较大等问题。而经典的非辅助学习(无监督)方法，如聚类分析、决策树等会将不同指标的重要程度视为等同的，忽略指标间权重的差异，极大地影响了分类精度。

以上的已有文献对于医保欺诈行为识别的算法进行了较深入的研究，得益于人工智能的快速发展，其研究大多数集中在辅助学习(有监督)方法，经典的非辅助学习(无监督)方法的相关文献还较少。基于已有文献的研究思路[6]，本文选择改进经典的非辅助学习－聚类分析，定义信息熵并以熵权法引入指标权

重，通过引入“依托权重改进的欧氏距离”，对不同指标间的“距离”进行改进，得到一种全新的“距离”用于层次聚类分析，聚类得到的孤立点将作为医保欺诈个案。

2. 模型准备

2.1. 数据的获取

本文采用 2015 年深圳杯数学建模夏令营 A 题数据[7]。

2.2. 研究方法

本文研究的主要问题是在没有先验条件，即没有已知欺诈数据的情况下寻找出可能的欺诈记录。如果通过人工筛选给出部分欺诈数据作为学习材料，存在较强的主观性。为解决这样的问题，本文将采用非辅助学习(无监督学习)方法。通过比较分析，本文将采用聚类分析算法，对数据进行分类处理，认为出现的孤立点为疑似欺诈点。作为一个实际的聚类问题，各种指标对于距离函数贡献程度是不同的，在传统的聚类分析算法中，无论是使用欧氏距离还是马氏距离，都将所有指标赋予相同的权重。本文考虑给重要指标赋予较大权重，可以很大程度提高聚类效果。本文通过定义信息熵并引入熵权法在无先验条件的情况下给出各指标的权重，再使用基于属性权重的改进欧氏距离进行聚类分析，从而通过描述数据的相似程度，区分出具有异于常态的个案。

2.3. 研究假设

- 假设每个患者是否发生欺诈行为是相互独立的，不存在相互学习与模仿。
- 假设医保欺诈只体现在医药费上，与住院费、治疗费无关(数据中仅存在各类药物费用)。
- 假设姓名、身份证号、出生日期、性别、电话号码等显著无效识别指标不会对是否欺诈造成影响。
- 假设欺诈记录只出现在孤立点中[8] (欺诈行为的单次金额及拿药次数会与正常情况偏离较大)。

3. 建立识别模型

由于欺诈数据(单张处方钱数较高、一定时间内开药次数较多)都会与正常数据之间有较大差异，因此本文以聚类过程中出现的孤立点为疑似欺诈点。而传统聚类方法忽略了各因素的权重差异，本文引入考虑各项指标权重的欧氏距离作为衡量数据间差异的指标。为了得出一组只依赖基础交易数据而完全剔除主观性的权重值，本文引入信息熵的概念：指标的变异程度越小，所反应的信息量也就越少，其对应的权值应该更低。对于原样本数据集，结合现实医保欺诈行为案例，求出选取的五个指标的的值。基于以上分析，本节将建立识别模型并检验，见图 1。

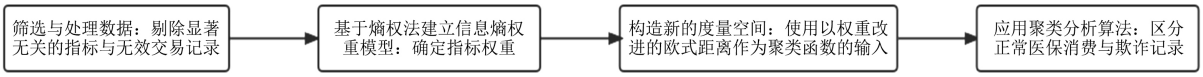

Figure 1. Flowchart for building recognition models
图 1. 建立识别模型步骤图

3.1. 分析及预处理数据

- **无意义数据的剔除及指标选取**

根据生活经验，可以排除掉身高、性别等冗余项的干扰，而对于所在省份等不齐全的信息也不能考虑对于结果的影响。

对于指标的选取：首先，欺诈行为的主体一定是有医保的患者。所以要在数据中筛选掉无医保的患

者的取药记录。

其次，林源、谌立平等人[9]指出：常见的两种欺诈类型表现为单张账单价格过高以及拿药次数过于频繁。因此单张账单的总价格和单个患者拿药次数应该分别作为判断是否为欺诈的第一项和第二项指标。

再次，考虑部分科室、或者个别医生可能存在不合理地开医嘱作为患者取药凭证的现象[10]，所以将执行科室以及开嘱医生 ID 作为影响判断是否欺诈的第三项和第四项指标。

最后，考虑到有一张欺诈类型的产生是刷他人的医保卡，可能存在部分人经常将自己的医保卡外借的情况[10]，所以将患者 ID 作为影响判断是否为欺诈的结果的第五项指标。

- **对原始数据的处理及结果**

从获取到的数据中[7]，导入患者 ID 和医保手册号——根据是否有医保手册判断患者是否有医保，根据判断结果忽略数据中无医保患者的购药记录。

为了得到单张账单的总价以及患者的购药次数，从数据中导入账单号、价格、患者 ID——将账单号相同的记录合并，进而计算出单张单据的总价；根据患者 ID 来统计每位患者的购药次数。

考虑到影响判断是否欺诈的其他因素，还应将数据中的执行科室和开嘱医生 ID 导入 Matlab 中。

- **数据处理后的形式**

处理后的数据应该包括六项参数——单张账单总价、患者拿药次数、患者 ID、执行科室、开嘱医生 ID、账单号。其中前五项参数为影响判断结果是否是欺诈的判断指标，第六项数据账单号为索引值，用于根据判断的结果检索该条记录的其他信息。

3.2. 熵权法确定权重

- **信息熵的引入**

信息理论的鼻祖之一 Claude. E. Shannon 把信息熵定义为离散随机事件的出现概率[11]。所谓信息熵，不妨理解成某种特定信息出现的概率。从信息传播的角度来看，当某种信息出现的概率较高的时候，表明它被传播得更远，或者说，这个信息被引用的程度更高，本文认为其占比权值相应更高。定义事件 X 的信息熵为：

$$H(X)=\sum_{i=1}^{n}\left[p(x_i)I(x_i)\right]=-\sum_{i=1}^{n}\left[p(x_i)\ln\left(p(x_i)\right)\right]$$

- **计算熵权**

首先进行数据的标准化处理，形成标准至非负区间，n 个评价对象，m 个评价指标构成的标准化后的正向化矩阵(如购药次数，次数越多，代表更可能发生医保欺诈行为，已经为正向化数据则不作处理)。

计算第 j 项指标下第 i 个样本所占的比重，并将其看作相对熵计算中用到的概率。

计算每个指标 j 的信息熵 e_j 及信息效用值 d_j，最后将其归一化得到熵权 W_j，步骤见图 2。

五个指标计算熵权后得到的结果为 $W=[0.1274,0.0662,0.4245,0.1874,0.1965]$。

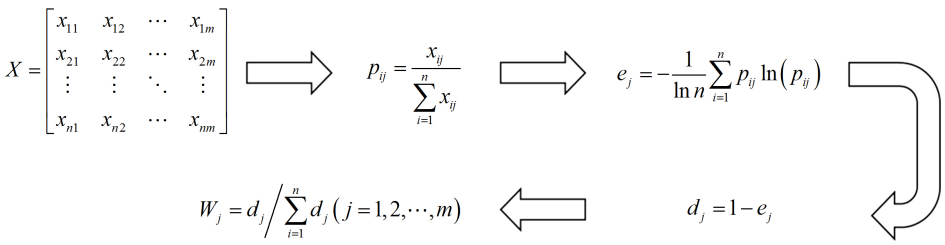

Figure 2. Diagram of the process of calculating entropy weights
图 2. 计算熵权过程图

3.3. 引入改进的欧氏距离

设 $X=\{x_1,x_2,x_3,\cdots,x_n\}m$ 为待聚类的医保消费记录数据集，每个数据 $x_i\,(1\le i\le n)$ 由个指标组成，即维数为 m，$x_i=\{x_{i1},x_{i2},x_{i3},\cdots,x_{im}\}$，其中 x_{ij} 是 x_i 的第 j 维属性。

为了描述样本点数据 x_p 和 x_q 在数据集 X 上的距离，通常使用的欧氏距离定义为

$$d_{pq}=\sqrt{\sum_{k=1}^{m}\left(x_{pk}-x_{qk}\right)^2}$$

上式中的欧氏距离将所有属性(指标)赋予了相同的权重，未考虑不同指标对于在医保欺诈识别中贡献程度的差异，以此种欧氏距离为分类依据的聚类分析的结果可能与实际情况产生较大差异。为改善聚类效果，本文引入基于属性权重的欧氏距离 $d_{pq}^{(W)}$，称之为改进的欧氏距离，其定义为：

$$d_{pq}^{(W)}=d_{pq}=\sqrt{\sum_{k=1}^{m}W_k\left(x_{pk}-x_{qk}\right)^2}$$

3.4. 依托权重进行层次聚类

- **层次聚类的定义**

给出各个指标的权重，通过对距离赋权的到新的距离定义，利用距离矩阵构建聚类树，而后可从聚类树中选择聚类树。

- **层次聚类数据准备**

首先对原数据进行标准化处理，取消量纲的影响，并大量减少计算复杂度(数据较大)。以对称矩阵的形式存储两个点之间的距离(平方欧氏距离)。

- **以给定距离权重构建聚类树**

由于数据量过大，电脑的内存有限，我们将 87,348 个样本每 10,000 个分别讨论。以下以第 1 至第 10,000 个样本举例。

使用 Matlab 编程求解引入 3.3 中定义的加入权重的欧式距离构成新的对称矩阵，此时我们得到了新的加入权重后的两两距离矩阵。使用 Matlab 内置层次聚类函数 linkage 构建决策树。

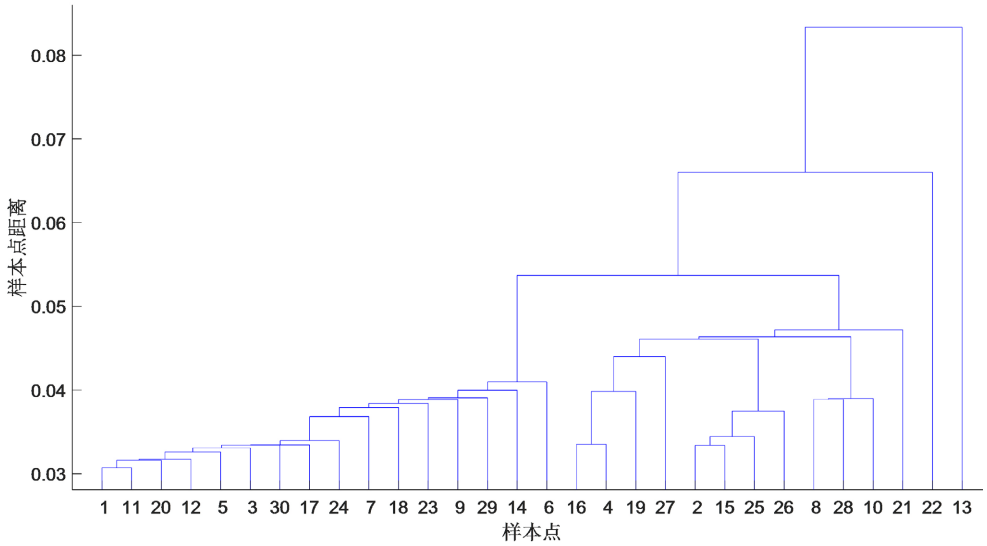

Figure 3. Icicle plot by dendrogram function
图 3. Dendrogram 函数绘制的冰柱图

- **结果处理**

使用 Matlab 内置函数 cluster 对已经构建的决策树提取一个聚类数为 3 的聚类。并使用内置函数 dendrogram 绘制冰柱图(由于节点限制，只能绘制部分)。

最后使用 Matlab 编程统计每一个聚类类目下所包含的记录数，以上操作均在程序中实现。第 1~10,000 个样本所得结果绘制冰柱图如图 3 所示。

3.5. 层次聚类结果

使用 Matlab 编程统计每一个聚类类目下所包含的记录数，第 1~10,000 的样本所得结果见表 1。

Table 1. Clustering results for selected samples
表 1. 部分样本的聚类结果表格

聚类类目	样本个数
1	2
2	9997
3	1

不难看出，聚类结果满足：欺诈数据(单张处方钱数较高、一定时间内开药次数较多)都会与正常数据之间有较大差异的条件。且完全的呈现孤立点分布。但聚类数是人为规定的，有一定的主观性，下将改变聚类数，探索规律并尽可能的多发现医保欺诈行为，聚类数为 3，4，5，6，7 的聚类结果见表 2。

Table 2. Clustering results with different number of clusters
表 2. 不同聚类数下的聚类结果

聚类类目	聚类数 3	聚类数 4	聚类数 5	聚类数 6	聚类数 7
1	2	2155	11	1	60
2	9997	7842	7831	7830	2095
3	1	2	2155	11	1
4		1	2	2155	7830
5			1	2	11
6				1	2
7					1

从结果中，我们发现当聚类数目为 4 时，疑似发生医保欺诈行为的样本数为 3，但聚类数为 5 时，样本数为 14，我们认为其不再满足条件所设：完全的呈孤立点分布。故我们选定聚类数为 4。使用对余下 7 万余个数据进行同样的处理得到的全部样本聚类结果见表 3。

Table 3. Results for all samples with a clustering number of 4
表 3. 聚类数为 4 的全部样本结果

聚类类目	1 至 10,000	10,001 至 20,000	20,001 至 30,000	30,001 至 40,000	40,001 至 50,000	50,001 至 60,000	60,001 至 70,000	70,001 至 87,348
1	2155	1	3	1	1	3	1	2
2	7842	9997	9994	9996	1	1753	9994	17,335
3	2	1	1	1	2	8243	2	8
4	1	1	2	2	9996	1	3	3

综上我们在 87,348 个个案中识别出发生医保欺诈行为的个案 43 个。
被识别为发生医保欺诈行为的个案见表 4 (仅展示前 20 个)。

Table 4. Cases identified as having committed health insurance fraud
表 4. 被识别为发生医保欺诈行为的个案

账单号	执行科室	病人 ID	总价	开嘱医生 ID	购药次数
5042603	173	163696	64.52	1180	27
5042604	173	163696	16	1180	27
5101762	329	656502	1315.21	2928	4
5101762	329	656502	1315.21	2928	4
5117504	329	658656	714.67	2928	11
5130775	210	189268	2489.07	1028	1
5116624	329	658656	713.98	2928	11
5116940	329	658656	714.67	2928	11
5145438	525	661781	831.45	2922	6
5140624	161	186919	1930.25	1060	7
5140636	161	186919	1930.25	1060	7
5217735	15	608684	1927.22	108	2
5235432	525	669778	1349.29	2922	7
5261516	525	198369	413.68	2922	6
5252243	525	239773	367.57	2922	8
5252258	525	239773	364.23	2922	8
5379295	210	233446	1179.39	1028	11
5379184	210	233446	1179.39	1028	11
5379206	210	233446	1179.39	1028	11
5383960	525	686465	597.33	2922	6

4. 结果分析

将由模型得出的可能欺诈记录进行人工复查，结果发现所得出的43条数据较好地被归属为两种情况，一是单张处方药费特别高，二是一张医保卡在数据给出的一个月时间内多次拿药。与现实中发生医保欺诈行为的案例十分吻合[8] [9] [10]，可以认为模型及结果是正确有效的。

参考文献

[1] 孙梦秋. 医保诈骗犯罪研究[D]: [硕士学位论文]. 扬州: 扬州大学, 2020. https://doi.org/10.27441/d.cnki.gyzdu.2020.000657

[2] 狄萱. 基于孤立森林和随机森林的医保欺诈识别系统[D]: [硕士学位论文]. 南京: 南京邮电大学, 2021. https://doi.org/10.27251/d.cnki.gnjdc.2021.000436

[3] 刘鋆. 医保欺诈数据异常深度学习算法分析研究[D]: [硕士学位论文]. 成都: 成都理工大学, 2020. https://doi.org/10.26986/d.cnki.gcdlc.2020.001013

[4] 李金灿, 徐珂琳, 於州, 魏艳, 仇春涓, 秦国友, 汪荣明, 徐望红. 大数据技术在医保反欺诈中的应用[J]. 中国医疗保险, 2021(1): 48-52.

[5] 陈富秋, 吕亚兰. 基于医保大数据的异常行为检测[J]. 技术与市场, 2021, 28(2): 18-20.

[6] 武优西, 侯丹丹, 李建满, 米少华.属性权重聚类算法的研究[J]. 小型微型计算机系统, 2012, 33(3): 651-654.

[7] 2015年深圳杯数学建模夏令营A题及其附件[Z/OL]. http://www.mcm.edu.cn/html_cn/node/41a9e7cecbbf3df4094aca089c10e2fd.html, 2015-04-14.

[8] 杜倩, 刘鸿宇, 胡琦. 社会医疗保险基金欺诈行为的扎根理论研究——基于58个医保欺诈刑事案件分析[J]. 法制与经济, 2020(8): 70-71+75.

[9] 林源, 谌立平, 宋曙光. 城乡居民医疗保险欺诈损失实证研究[J]. 怀化学院学报, 2020(4): 50-54.

[10] 姚强, 杨菲, 郭冰清. 基本医疗保险“欺诈骗保”现象的影响因素及路径研究——基于我国31个省级案例的清晰集定性比较分析[J]. 中国卫生政策研究, 2020, 13(11): 24-31.

[11] Shannon, C.E. (1956) The Zero-Error Capacity of a Noisy Channel. *IEEE Transactions on Information Theory*, **12**, 8-19. https://doi.org/10.1109/TIT.1956.1056798