基于数据挖掘的医保欺诈主动发现

电机系 2013010932 方诗卉
电机系 2013010935 谭振飞
电机系 2013010946 贾　鑫

摘要

目前社会上存在着一些不法分子在履行参保缴费义务上虚构事实，隐瞒真相，以骗取医保权益，或在医疗行为上虚构事实，隐瞒真相，以骗取医保基金或医保待遇。这类欺诈行为在各个国家普遍存在。这些违法行为已经给我们国家带来了极大的经济损失，严重影响我国医疗行业的进一步发展。而在海量就医病患信息中识别出骗保行为，靠人工的力量显然是不行的。为了能高效地初步识别出骗保嫌疑对象，本文使用聚类分析、异常点挖掘、残差分析等算法对病患就诊信息进行数据挖掘。针对不同的骗保手段，给出了四个骗保识别因子作为评判标准，并由此通过一种平方平均的方法得到了一个集中的骗保嫌疑评判指标。通过计算每一例就医记录的嫌疑指标，并与设定的阈值比较，可以初步锁定骗保嫌疑对象。再在这些对象中进行更精细的人工调查，便可有效地识别骗保行为。

关键词：欺诈识别、数据挖掘、异类识别、SNN 相似度、识别因子

一、问题重述

医疗保险欺诈，是指公民、法人或者其他组织在参加医疗保险、缴纳医疗保险费、享受医疗保险待遇过程中，故意捏造事实、弄虚作假、隐瞒真实情况等造成医疗保险基金损失的行为。骗保人进行医保欺诈时通常使用的手段，一是拿着别人的医保卡配药，二是在不同的医院和医生处重复配药。下面这些情况都有可能是医保欺诈：单张处方药费特别高，一张卡在一定时间内反复多次拿药等。根据附件中 6 个表格（病人资料、费用明细表、医嘱表、医嘱子类、核算分类、患者类别）中的数据，找出可能的欺诈记录。

二、问题分析

关于医疗保险欺诈，骗保人会在医保的使用过程中留下痕迹，会出现非正常消费的情况。在本问题中，需要通过检索分析消费记录来进行选择。

首先，根据表格 2.1——病人资料，可以看出，有一部分人在医院的记录中使用了医保卡，而另一部分人没有使用。所以对没有使用医保卡的情况可以断定出没有骗保嫌疑。同时还能看出一些其他问题，比如病人在医院死亡，对这样的病人也可以基本断定不会有骗保嫌疑。所以在处理数据的第一步需要对原始数据做预处理，去掉无骗保嫌疑的人员，后续的数据处理只针对剩下的人。

对于剩下的无法直接排除嫌疑的人，不能一概而论，由于骗保行为多种多样，所以在医保消费的过程中也存在不同的特点。这些特点大体上可以分为以下三种：

1. 数量特点：由于骗保的人需要获得更多的利润，所以他需要在通过医保卡购买很多药，来实现骗保的行为。而这一点在表格中的体现便是某一个医保卡号对应的消费量多于正常值。所以在这一方面可通过找出每一个 ID 号所对应的消费量，并与均值相比较，来刻画其骗保的可能性。
2. 频次特点：对于骗保人来说，有可能为避嫌而分多次购买药品。这一点在表格中的体现为某一个医保卡实现无规律的多次消费。所以在这方面可以通过找出每一个 ID 号对应的消费频次，其中的异常情况既代表着骗保出现的可能情况。
3. 类型特点：根据表格 2.3——医嘱项，可以分析出，对于一个的医嘱项，在每个使用的病人上应该具有大致相同的量，而一旦有一个量“与众不同”，既代表着他有成为骗保人的可能。所以在这方面可以通过刻画这个大致相同的量（即例均费用）来比对每一个的医保使用量，从而刻画出骗保的可能性。

通过对这三个方向数据的综合分析，就可以基本刻画出骗保可能性的大小了。为了模型更精确更具普遍性，可以继续考虑其他的影响程度比较小的因素，比如使用医保卡但并未使用身份证的患者有一定的嫌疑，等等。

三、模型假设及符号说明

1. 模型假设

我们建立的数学模型，是针对于每一个人，对每一个人在不同的指标上进行评分，每一

种指标 0 作为嫌疑最小，1 作为嫌疑最大，通过这样的多个识别因子，在空间构建一组多维向量，而在这个多维空间中，距离大多群体最远的个体的骗保嫌疑最大，可以表示出当该人的值偏离远点到一定范围之后就可以视作骗保。

此外，构建模型分析之前我们进行以下假设：

1.1 假设不用医保卡消费的患者不存在骗保费用；

1.2 假设一次开的每一种药量是一个定值，当某个病人开的药偏离这个标准值越远，他的骗保嫌疑就越大；

1.3 假设有一些药是需要按疗程服用的，这样同一个病人间隔一确定时间（如 7 天）来买同一种药视为比较正常，而间隔时间较短或者频率不稳定的骗保嫌疑较高。（例均费用）；

1.4 假设对于不同医嘱表和医嘱子类有着比较明显的差距，医嘱表存在不会骗保的项目，同时在医嘱表的一些不会骗保的项目中有消费记录的病人也基本排除骗保嫌疑；

1.5 假设存在某一个人使用不同的医保卡在同一时间重复买药，这样将这些医保卡相互关联操作。同时提高了骗保的嫌疑。

2. 符号说明

符号	含义
y_k	患者年龄；
c_k	就医消费平均值；
$\overline{\delta_1}$	每一患者的标准偏差；
q_1	识别因子；
N_k	偏差分布区间中点；
P_k	怀疑对象所占频率；
α、β	权重系数因子
U	就诊病例五种特征类型所构成的矩阵；
x_{ij}	矩阵中的某个量；
x_i	矩阵的 i 行平均值；
S_j	矩阵的 j 列的方差；
m	取值的富豪特性种类；
D	距离矩阵；
q_2	识别因子；
d_{out}	离群点距离最近聚类中心的距离；
$\overline{d_n}$	离群点距离其 k 个近邻的平均距离；
A	归一化因子；
H_I、α_i	各识别因子的加总权重；
Q_i	嫌疑识别因子。

四、模型的建立与求解

0. 数据预处理

通过表格 2.1 可以看出，在所有的病人就诊记录中，有一部分有医保卡号记录（即使用了医保卡），有一部分没有医保卡号记录（即未使用医保卡）。由于未使用医保卡的病人在骗保方面的嫌疑为 0，所以在数据处理中需要先剔除掉未使用医保卡的情况。这样既减少了误判的可能，又避免了重复计算加快了运算效率。

1. 费用额度异常筛选

在进行医保欺诈的识别筛选中，最直接的指标就是费用额度。此部分将从每人消费总额与每单消费总额两方面确定识别因子。

1.1 基于年龄分布拟合的残差异常识别

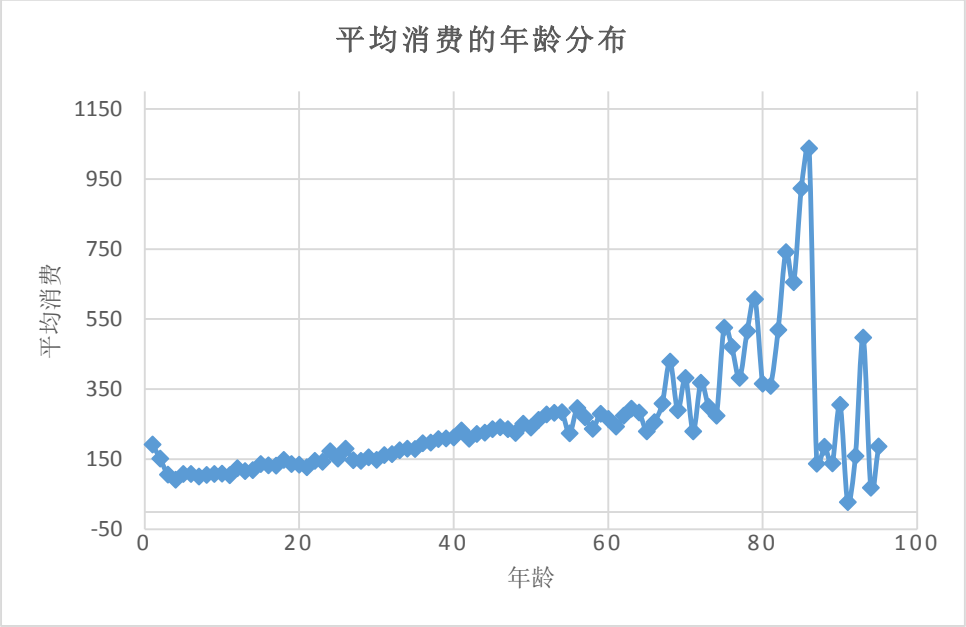
不考虑个体差异，则每人一定时期内就医开支和年龄存在着较大的相关性。直观上理解，年龄较小或较大者身体状况较差，其一定时期内就医开支也将高于青壮年人群。基于此分析，找出平均费用额度对年龄段的分布情况，对该分布进行回归分析，求出拟合函数。之后以该函数作为参考标准，对每个人的消费数据进行检验，求出每人消费额的残差，通过残差大小确定其该项指标的识别因子。

1.1.1 费用年龄分布拟合

所有年龄为y_k的患者，他们的就医消费平均值为c_k。c_k与y_k的函数关系为：

$$c_k = F(y_k)$$

将《2.2.1 病人就医分类汇总表》与《2.1 病人资料表》按 PAPMI_ID 连接，之后按年龄以平均的方式进行分类汇总。得到$c_k - y_k$分布图：



可以看出该分布规律和预期基本一致，医疗开支基本随着年龄的增大而增加，幼儿和老年人的开支有显著增加。但也不难发现 65 岁之后的分布波动剧烈，这是由于年龄较大的人数较少，统计的结果受个体影响较大。

对表中年龄为 4 至 60 岁的人群的平均消费分布进行最小二乘拟合，其余年龄段使用平均值作为对应的标准函数值。

1.1.2 残差分布分析

计算每个患者的总就医消费与他所处年龄段的标准参考函数的偏差。由于不同年龄的消费数额不同，因此需要对偏差进行归一化。计算每一患者的标准偏差：

$$\overline{\delta_1} = \frac{c_k - \overline{c_k}}{\overline{c_k}}$$

得到标准偏差的频数分布表如下：

区间	[−5, 0]	[0, 5]	[5, 10]	[10, 15]	[15, 20]	[20, 25]	[25, 30]	[30, 35]
频数	24752	10717	313	35	5	10	4	2
质疑	×	×	0	0	0	0	0	0
频率	/	/	0.848	0.0949	0.0136	0.0271	0.0108	0.0054

由此表可以看出，大部分数据的标准偏差分布集中在[-5,5]范围内，因此可以认为不在此范围内的患者在就医开支金额这一识别指标上就有一定嫌疑。

1.2 识别因子确定

虽然已经得到标准偏差的分布状况，具体确定识别因子q_1的取值有一定困难。但是我们认为，基于单人消费总额分布异常的识别因子q_1至少应满足如下条件：

a) 识别因子的取值范围：[0,1]；

b) 标准偏差越大，识别因子q_1的值越大；

c) 相应区间的频数越少，识别因子q_1的值越大；

d) 有明显较多数据的区间可直接排除其嫌疑。

根据这些基本条件，可以按如下公式确定识别因子：

$$q_1 = \frac{N_k^{\alpha}}{P_k^{\beta}}$$

其中：

N_k——表示偏差分布区间中点

P_k——怀疑对象所占频率

α、β——权重系数因子

这一关系近似表示了识别因子q_1与决定因素的关系，但权重系数因子α、β需要在最后的识别因子加总时进行标定。

2. 药品种类与数量异常挖掘

同一类药品或治疗手段，它们的单次使用数量、适用人群、疗程周期等应具有一定相似

性。而骗保行为往往会违背这些常规，因此这也可以作为识别骗保行为的一个指标。下面将使用基于 SNN 相似度的异常点筛检算法进行识别分析，同时借鉴 MACLU 聚类异常挖掘算法中高维混合数据的处理方法对算法进行改进。

2.1 数据预处理

本问题需要考虑药品种类（医嘱子类）、医嘱类别、单次使用数量、就诊人性别、疗程周期、五个不同的指标，它们采用不同的度量单位，数值差别可能很大，会造成 SNN 相似度计算时距离计算得较大偏差。同时，药品种类和医嘱类别属于符号型数据，而后三者为数值型数据，即这五个维度的数据包含混合数据类型，因此不能直接进行聚类挖掘计算。因此需要首先对数据进行标准化、正规化、符号类型数值域映射处理。

2.1.1 数据标准化

每一就诊病例的上述五种特征类型数据构成矩阵 U，每一病例的五种特征数据构成矩阵的一个列向量。标准化需要使得各特征类型的数据均值为 0，方差为 1.处理过程为：

$$x'_{ij} = \frac{x_{ij} - \overline{x_\iota}}{S_j} \quad (i = 1,\cdots,5; j = 1,\cdots,n)$$

其中：

$$\overline{x_\iota} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$\mathrm{S_i} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_\iota}\right)^2}$$

数据经以上标准化之后数据范围很可能不在[0,1]上，还需进行下面的正规化变换。

2.1.2 数据正规化

正规化变化处理如下：

$$x'_{ij} = \frac{x_{ij} - \min\limits_{1\le i\le n} x_{ij}}{\max\limits_{1\le i\le n}(x_{ij} - \min\limits_{1\le i\le n} x_{ij})}$$

其中变化公式的分母是数据矩阵 U 第 i 行的极差。经此步变换后各变量最小值为 0，极差均为 1，并且各特征属性的基点相同、波动范围一致，这样就方便后面的计算分析。

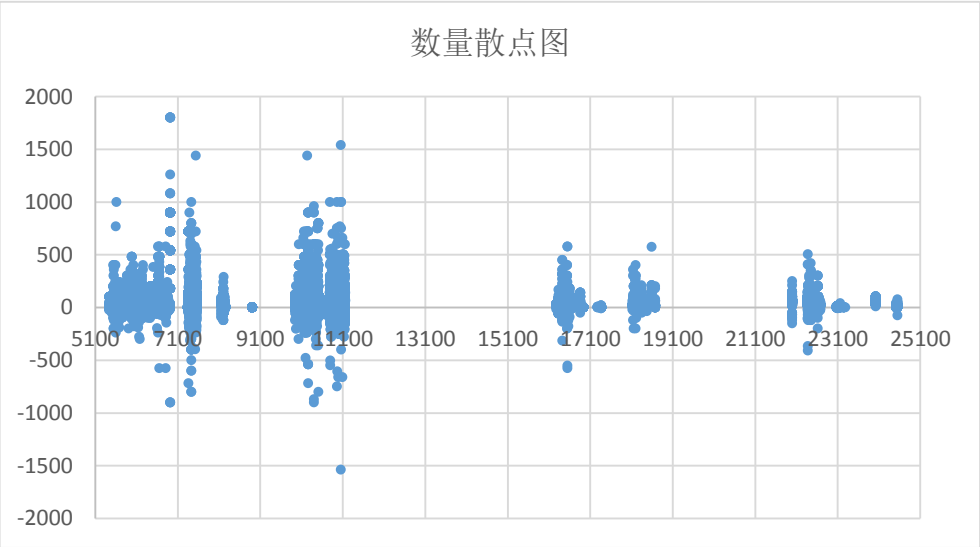
2.1.3 符号类型特征向数值域映射

为了使得符号类型特征能像数值特征一样进行处理，此步采用如下编码映射方法进行处理：

为了保留符号类型特征的等同行本质，映射具体做法是：对于有 m 种不同取值的符号特性，用 m 位二进制编码对其编码，当且仅当特征取值为第 j 种值时，其编码中

的第 j 为为 1，其余为 0.

但具体到本问题，由于包含的药品类别十分之多，就有可能使得特征类型的像矢量过大。因此直接作出以药品类别的代码与该药品（治疗）的使用数量的散点图进行观察：



可以看出虽然药品种类庞杂，但一段时间内只有不到一半被患者用到，而其中又有相当一部分的分布较为集中。可以认为只有使用数量分布分散的药品种类才存在骗保的可能性，因此实际需要编码映射变化的药品种类是可以承受的。事实上，还可以对映射后的像矢量进行 PCA 主成分分析进行降维处理。具体实现太过复杂，此处仅提出概念，不做进一步讨论。

2.2 SNN 相似度异常点筛检

将数据进行合适的预处理之后，下面计算各样本点之间的相似度矩阵，根据相似度矩阵确定各数据点之间的连接关系，确定出聚类中心和异常点数据，进而筛检出异常数据。具体实现步骤如下：

2.2.1 构造距离矩阵

计算 U 矩阵中各 5 维列向量之间的欧式距离，构造距离矩阵 D。其中距离计算公式如下：

$$d_{ij} = \sqrt{\sum_{k=1}^{8}(x_{ki} - x_{kj})^{\wedge}2}$$

值得说明的是，距离矩阵 D 是对称阵。

2.2.2 建立 k 近邻列表

对于每个样本点，在 D 中找出其最近的 k 个近邻，将其存入 k 近邻表 KNN 中，其中 KNN(i)表示第 i 个样本点的 k 近邻列表。

2.2.3 计算SNN 相似度

在 SNN 表中计算两两互为最近邻居的样本点见的 SNN 相似度。SNN 相似的定义为：建立相近节点 i 和 j 之间的连接，则每个连接之间包含两个样本的 s 个共同邻居，称 s 为这一连接的连接强度，两个最近节点之间的连接强度即为 SNN 相似度。

计算得两两样本点之间的 SNN 相似度，将其存入 SNN 相似度矩阵 S。

2.2.4 检测异常点

基于 SNN 相似度矩阵 S，确定检索阈值 t，是要 SNN 相似度大于 t，则建立两个样本点之间的关联，将其聚为一类；最终没有和任何点聚为一类的样本点即被认为是异常点。

2.3 识别因子q_2确定

在上一步的 SNN 相似度异常点挖掘中，找出了明显没有聚类的数据点，下面给出识别因子的确定方法：

首先对于聚类数据，其识别因子q_2取 0；

对于离群点，其识别因子：

$$q_2 = A \cdot \frac{d_{out}^{\gamma}}{\overline{d_n}^{\sigma}}$$

其中：

d_{out}——离群点距离最近聚类中心的距离

$\overline{d_n}$——离群点距离其 k 个近邻的平均距离

A——归一化因子

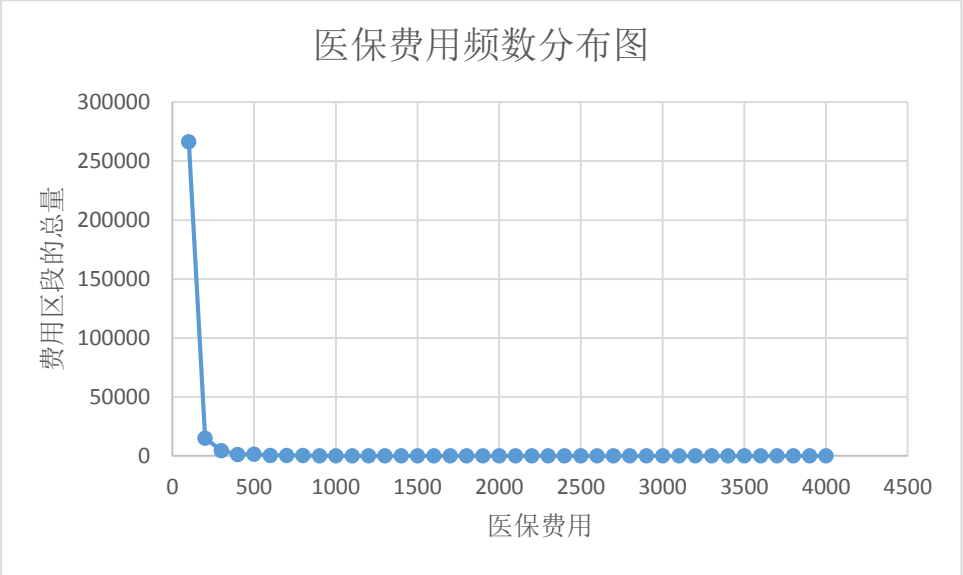
3. 一单大额的异常识别

在骗保的过程中，由于不是真正需要治病，而只是单纯的用医保购药，所以可能会一笔使用医保卡消费大额数据，从而在医疗记录中留下痕迹。

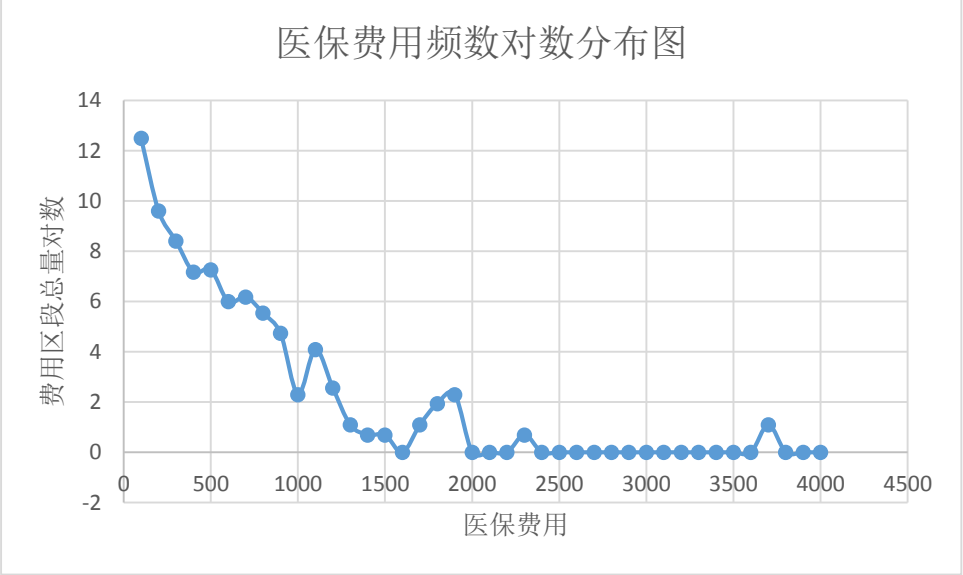
对于消费记录来说，由于总量很多，同时绝大多数消费额度较低，应当大致满足指数分布，即随着消费额的增加，消费人数大体呈指数衰减，而对于偏离指数分部的数据即存在骗保嫌疑，偏离越多嫌疑越大。

3.1 数据分析

对于消费总量一栏数据进行处理，考察从 0 到 4000，每 100 为一档中的数据，绘制频数分布直方图，看每一个频带中落入数据的个数，这些数据也基本上保持指数分布，绘制图线如图所示：



从上图中可以看出，绝大多数的数据分布在小于 100 的区间，数据近似呈指数分布。对指数分布的各个数据取对数，得到一个近似线性的直线，如下图所示：



从本图中可以看出，在费用小于 1500 的区间基本呈直线下降，拟合得到直线的相关系数为 0.9686。从这个分布可以看出，当一单医保费用等于 1500 时，频次基本降为 0。所以说可以定出阈值为 1500。当一单费用超过 1500 时视为非正常消费，存在骗保嫌疑。

3.2 识别因子确定

由于之前的部分在计算每个医保号的消费总额时在一定程度上包含了一单大额的数据信息，所以本模型只是上一个模型的拓展，和单人消费总额模型共同构成本问题的第一个维度，可以由此确定第三个识别因子q_3

经选取，可以得出最终有病历号为：217527，267254，387776，397488，405032，463011，477945，539869，579502，589176，608684，612657，615989，

628287，639799，660150 的符合这一点情况，其中，267254，387776，477945 这三个的消费记录和别的没有相同的项目，有很大的骗保嫌疑，剩下的几个都有和其他人在不同时间的相同消费记录，骗保嫌疑较小。

4. 辅助识别

4.1 频次特点的刻画

由于有一些药品或者治疗方案是讲求疗程和阶段的，所以在有些的医保消费记录中，存在同一张医保卡，每隔一确定时间，重复购买同一种药品的情况。而对于骗保者来说，一般不会隔一确定周期来购同一种药，尤其可能会在短时间内多次购买同一种药。

对于本模型来说，考虑其消费日期记录分部周期性较好的病人骗保的概率较低，而周期性较差，尤其是在短时间内多次购买同一药品的，骗保的概率较高。

在实际计算的过程中，需要跟踪同一个人买的同一种药，计算买药周期，比对周期偏差，对于周期偏差的大小赋予不同的数值，来刻画这个维度上的骗保嫌疑。

4.2 实名制的可信度刻画

在病人资料一栏中，有很多病人的有关信息，比如身份证号，病人姓名等等。这些项目只有一部分的记录中有，在一定程度上刻画出了实名制可信度情况的刻画，一般的骗保人很少会使用实名制记录。

对于本模型来说，是要大大减少实名制登记的病人骗保的概率，可以通过提高其他未实名制登记的人的骗保概率来做到相对减少其相对值的大小。

4.3 多个医保共同使用的情况刻画

对有些骗保人来说，他有可能在很短的一段时间内用很多张医保卡进行消费，以做到骗保的目的，同时不易察觉。所以需要针对这一问题使用特殊的模型进行甄别区分。

对于本模型来说，应当将在间隔时间较近买进同一种药的每一个人做以关联，在前面的分析基础上，让每个相互关联的卡有一个统一的关联系数，在一个不信任度增加的时候，其他的也有所增加。这样可以在一定程度上甄别这一问题。

4.4 特殊医疗使用的刻画

对于骗保人来说，更多的会使用，药品或者注射针剂进行骗保，而对一些像疫苗、或者大仪器的使用方面，一般不会出现骗保的情况。

对于本模型来说，是通过表格 2.5 中的医疗项目分类，将基本不会骗保的项目取出，索引使用过这些项目的病人，然后大大降低这些病人的骗保概率。

5. 识别因子的归总

通过前面的就诊开支金额异常筛选模型、大单异常筛选模型、账单特征属性异常点挖掘

模型和其他辅助识别模型，我们得到了若干互相独立的识别因子。这些识别因子分别刻画了不同的骗保手段，为了使用一个更集中的指标识别骗保行为，还需对这些分立的识别因子进行合理的组合，得到一个最终的识别指标。

考虑到前三种异常中居其一即可认为有较大骗保嫌疑，因此对前三个模型得到的识别因子进行平方和加总；而对于辅助识别，可以看做是共同决定了一个新的识别因子，将该因子参与前面的平方和加总。最终得到的嫌疑识别因子计算如下：

$$Q_i = \sqrt{H_1 q_1^2 + H_2 q_2^2 + H_3 q_3^2 + \frac{H_4(\alpha_1 q_{41} + \alpha_2 q_{42} + \alpha_3 q_{43} + \alpha_4 q_{44})}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4}}$$

其中：

H_I、α_i为各识别因子的加总权重，在实际中可以根据各类骗保行为的发生概率、不同骗保行为的检测灵敏度需求和经验进行选取。

按照实际需求选定 Q 的阈值Q_t ，对每一个病例计算其嫌疑识别因子并与Q_t比较。若其超过识别阈值则可认为其有较大嫌疑，接下来可以进行更具体的人工调查。

五、模型的评价与推广

在我们建立的数学模型中，首先通过一系列小模型刻画每一个具体的问题，这样使得模型更加准确，与问题贴合更紧密；其次通过一共三级的大模型进行整理汇总，可以在这些问题中体现出分清主次的特点，对主要问题进行重点描述与分析，将每一个小模型进行整合，或是加权平均，或是求平方和的办法，能够突出主要部分，并且通过不同的权重调整各个部分对最终结果的影响，在实际的计算过程中能够做到灵活多变。

本模型同样可用来检查其他保险的一些骗保问题，同时也可以用来检查偷税漏税的问题，还有考试中的代考问题，以及在各种大问题中的虚假信息，比如大公司甄别下属部门的真实工作与虚假工作的信息。能过在很多方向上得到推广。

参考文献

[1] 夏宏,汪凯,张守春.医疗保险中的欺诈与反欺诈问题[J].现代预防医学. 2007 年第 34 卷第 20 期：3907，2007.

[2] 姚奕,孙祁祥,林山君,范庆祝. 医疗保险欺诈识别——基于中国医疗费用商业保险的实证研究.北京大学经济学院工作论文[D]. C-2014-003.2014 年 3 月 25 日.

[3] Pang-Ning Tan, Michael Steinbach, Vipin Kumar 著.范明,范宏建译.数据挖掘导论(完整版)[M].北京:人民邮电出版社,2011.

[4] Jiawei Han, Micheline Kamber 著.范明,孟小峰译.数据挖掘概念与技术[M].北京:机械工业出版社,2008.

[5] 梁爱琴.数据挖掘关联算法在医保系统中的应用[D].北京:北京工业大学,2008.

[6] 田青华.基于多 Agent 医疗欺诈行为检测系统的研究与设计[D].浙江：江苏大学,2009.

[7] 潘芳.基于贝叶斯的防病患欺诈模型研究[J].现代商贸工业.2014 年第 10 期：80-82,2014.

[8] 牛晓辉.新农合住院费用的分析及异常值筛检方法研究[D].湖北：华中科技大学,2012.