计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(6):34−39 [doi: 10.15888/j.cnki.csa.006363]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 BP 神经网络的医保欺诈识别①

刘　崇, 祝锡永

(浙江理工大学 经济管理学院, 杭州 310018)
通讯作者: 刘　崇, E-mail: mxlclc@163.com

摘　要: 医疗保险欺诈是指在参加医保的过程中, 通过故意捏造、虚构事实等方法骗取医保基金或医保待遇, 造成医疗保险基金损失的行为. 有效地识别医保欺诈对医保基金的健康使用有重大意义. 本文运用 BP 神经网络实现医保欺诈的主动识别, 并利用 Logistic 回归分析对神经网络模型进行改进, 降低弱因子对神经网络识别的干扰. 此外, 应对欺诈数据的稀缺问题, 采用只取正常数据训练神经网络模拟函数曲线的模式. 实证表明, 该方法对医保欺诈具有较好的识别能力.

关键词: 欺诈识别; BP 神经网络; Logistic 回归分析; 模拟函数曲线模式

引用格式: 刘崇,祝锡永.基于 BP 神经网络的医保欺诈识别.计算机系统应用,2018,27(6):34–39. http://www.c-s-a.org.cn/1003-3254/6363.html

Medical Insurance Fraud Identification Based on BP Neural Network

LIU Chong, ZHU Xi-Yong

(School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Medical insurance fraud refers to the behavior of medical insurance fund or medical insurance coverage, which causes the loss of medical insurance fund through the method of deliberately fabricating and fictitious facts. Effective identification of health insurance fraud is of great significance to the rational use of health insurance funds. This study uses BP neural network to realize the active identification of health insurance fraud, and uses logistic regression analysis to improve the neural network model and reduce the interference of the weak factor to neural network identification. In addition, to deal with the scarce problem of fraudulent data, the model of neural network simulation function is used to train neural network. The empirical evidence shows that this method has better identification ability for health insurance fraud.

Key words: fraud identification; BP neural network; Logistic regression analysis; simulation function curve mode

我国医疗保险制度经过几十年的建设和改革, 覆盖人员总数已经超过了 13 亿. 然而自城镇职工基本医疗保险和新型农村合作医疗制度实施以来, 骗取医保基金的行为不断出现. 医保欺诈是欺诈相关人员通过参保或虚构、夸大保险伤害等方法骗取保险赔偿, 具有故意和有目的性、侵犯他人利益、严重的社会危害性等特征. 常见的欺诈手段有就医资格作假、病因作假、票据作假、处方作假、医疗文书作假等. 医保欺诈已对医保基金安全构成了重大的威胁, 严重妨碍了我国医疗保险政策的长期可持续发展.

由于医保欺诈中行为主体的复杂性和实施手段的隐蔽性, 使得欺诈行为很难通过直观判断发现. 但医保数据客观反映了参保人的行为, 对海量的医疗结算信息进行数据分析, 寻找医保欺诈潜在的行为规律, 识别其中的欺诈行为, 能有效避免医保基金损失. 旧有的基于数据统计方法的分析工具已经无法有效地违规数据

① 收稿时间: 2017-09-18; 修改时间: 2017-10-10; 采用时间: 2017-10-17; csa 在线出版时间: 2018-05-28

进行识别, 而数据挖掘和人工智能技术的发展为医疗保险欺诈识别提供了新的方法和途径. 运用数据挖掘和人工智能方法从海量的医保数据中发现潜在的违规信息, 是医保欺诈问题识别的重要研究方向.

1 研究现状

保险欺诈识别的研究根据方法和工具大致可以划分为三个阶段:

第一个阶段对保险欺诈识别的研究方法以统计回归为主, 核心是建立回归分析模型以查找关于欺诈问题的关键指标, 并赋予相应的权重, 以此实现对保险欺诈的识别与审核. 该阶段多采用二元离散选择模型作为进行欺诈识别: Artis, Ayuso 和 Guillen[1]采用 logit 离散模型对西班牙机动车保险索赔数据进行欺诈分析, 并针对数据值缺失的情况建立了改进的 AAG 模型统计, 得到众多业界同行的认可. 除了离散模型, Brockett, Derring[2]采用对比思想建立了 PRIDIT 模型, 用以研究欺诈概率; Pinquet [3]设计了消除偏差的两方程模型, 对比单方程模型, 该模型较好的消除了理赔样本选择偏差对审计结果的影响. 国内相关的实证研究文献较少, 叶明华[4,5]采用 logit 模型建立欺诈识别的指标体系, 借助平安保险的机动车索赔数据进行实证研究.

第二阶段是运用数据挖掘理论对保险欺诈进行研究. Viveros [6]使用关联规则和神经元分割对医疗保险数据进行分析, 用以发现未知行为模式; Chiu[7]用改进的 Apriori 算法构建了欺诈识别模型, 可以识别特定医疗服务中的疑似欺诈数据; 何俊华[8]基于垂直数据格式的频繁模式发掘, 设计了 CBM 和 MaxCBM 算法用于检测骗保行为; 陈亚琳[9]采用聚类分析和分类决策树算法建立预测模型, 用于识别某位病人在一段时间内是否存在保险欺诈行为; 唐暟宜[10]将主成分分析引入聚类分析, 结合两种方法对医保数据进行综合评价.

第三阶段是人工智能技术阶段. 人工智能技术的引入使得保险欺诈识别的研究取得重要突破, 尤其是神经网络技术的应用. 澳大利亚医疗保险委员会的 Hubick[11]首次在保险欺诈识别中应用 BP 神经网络, 为后来者提供了新的研究思路; Hawkins 等[12]用改进的三层 BP 神经网络改善了保险欺诈的识别精度; Lious 等[13]用神经网络、逻辑回归合分类树对台湾健康保险系统中数据进行实证研究, 实证表明 3 种方法在保险欺诈的识别中都有不错的表现, 其中神经网络方法识别能力最优; Maes[14]等分别采用神经网络和贝叶斯网络进行欺诈识别, 其中贝叶斯网络识别更为精准.

综上所述, 早期关于保险欺诈的研究集中在统计回归方法的应用上, 能在特征明显的数据上取得不错的识别效果. 但随着欺诈手段的复杂化、多样化、隐蔽化, 用该方法很难达成预期的识别效果, 尤其是基于单因子的识别方法局限更为明显. 数据发掘的方法应用上, 利用孤立点方法挖掘异常数据等方法为欺诈识别提供了有效的途径, 但数据挖掘方法是基于大量的原始资料的收集与分析, 从中提炼出有价值的信息, 在数据体量和分析成本上要求较高. 神经网络方法因具备自我调节能力, 使得模型更够适应新的数据和欺诈手段, 在保险欺诈识别上取得较为广泛的应用, 具有不错的借鉴意义, 但模型的训练受数据样本选择影响较大, 若样本训练不当, 难以达到预期效果.

考虑现实因素, 国外保险行业信息化、商业化程度较高, 较多研究成果已经取得了不错的应用. 而国内, 大部分医院虽已实现信息化建设, 但是彼此数据是独立的, 并未打通; 各省市医保信息也彼此独立、互不联网; 商业化保险的发展也存在不足. 又出于个人信息隐私和商业机密保护的考虑, 实际可获得的研究样本有限, 这一系列因素制约了国内关于医保欺诈实证研究的发展. 此外, 国内外医疗保障制度和社会经济环境也存在差异, 因此不能完全照搬国外的方法. 针对我国医疗保险的现状, 克服实证方法的缺乏和数据的限制, 构建一个稳定的欺诈识别模型对我国医疗保险事业健康发展有积极意义.

2 算法选择

本研究的主要目的是通过模型来识别用户是否存在欺诈行为, 被解释变量是一个布尔形变量, 本研究期望建立一个有监督的二分类模型用于欺诈识别.

医保欺诈涉及骗保人的心理、技术手段等众多因素及其相互作用, 具有高度非线性的特征. 这使得传统的利用指标加权重的传统分析方法如主成分分析、因子分析等很难获得较好的识别效果. 而 BP 神经网络在自适应性、自组织上表现突出, 具备良好的学习、容错及抗干扰能力, 自变量可以是连续的, 也可以是离散的, 可以识别复杂变量间的非线性关系, 因此选择 BP 神经网络进行医保欺诈识别具备可行性.

BP 神经网络是一种按照逆向传播算法训练的多层前馈神经网络, 由信号的正向传播与误差的反向传播两个过程组成. 正向传播的过程中, 输入样本从输入层传入, 经过多层隐层处理后, 传向输出层. 其中每一层神经元的状态只受相连的上一层神经元状态影响. 若网络的实际输出和期望输出不符, 则再进行误差的反向传播. 反向传播过程中, 误差信号按照原路逆向传播, 在传至隐层时, 对隐层各个神经元的权值进行修正, 使得最后的误差达到最小.

但考虑到基于经验分析选择的欺诈因子存在主观性, 在实际中可能不是判断欺诈与否的显著因素. 通过 Logistic 回归分析先行检验欺诈因子的有效性, 精炼解释变量以减少噪声的干扰, 或可以提高神经网络识别的精准度. 因此, 本文利用 BP 神经网络和 Logistic 回归分析方法的互补性和相互纠错性, 使之有效融合, 提出一种基于 BP 神经网络和 Logistic 回归分析结合的医疗保险欺诈识别方法. 经验证, 该方法能有效的识别医疗保险欺诈数据, 且结合回归模型剔除弱因子在一定程度上可以提升 BP 神经网络的精确度.

3 模型的搭建

3.1 欺诈因子确定

本文数据来源自深圳市南山、蛇口、西丽三家医院信息系统采集的病人和费用核算信息, 数据集中包含索赔人信息、索赔信息、治疗情况, 例如身份证号、医疗手册号、性别、购药单价、总价、医嘱类及医嘱子类等.

首先将性别作为医保欺诈识别的考虑因素. 有统计资料显示保险公司的欺诈案件中, 性别比存在较大差异: 性别导致不同的风险倾向, 男性冒险意愿强于女性. 此外, 年龄的不用往往也影响欺诈概率, 不同年龄层次的人在面对社会压力和疾病发生概率上都有差异, 因此将年龄也纳入欺诈识别的考虑因素之一.

再对比研究数据和欺诈方法, 可以发现医保欺诈行为反映在数据上有以下几个明显特征: 不满足医疗保险号与身份证的唯一识别、单张处方费用特别高、同一病人买药频率过高、在不同医院和医生处重复配药. 因此, 可以考虑将买药频率、买药总花费、买药总数量、医嘱项作为欺诈因子.

在某些情况下, 科室可以通过与参保人串谋伪造病历和票据, 以骗取医保基金, 因此可将科室作为欺诈与否判断的参考因素. 因研究数据涉及科室相关数据有下嘱科室、执行科室、病人科室三项, 在此都纳入欺诈因子的考虑中.

基于以上几点, 本研究设定了 9 个欺诈因子作为欺诈指标, 并通过 BP 神经网络来进行判定其设定的合理性, 9 个神经网络输入因子分别为性别、年龄、买药品频率、买药总花费、买药总数量、下嘱科室、执行科室、病人科室和医嘱重复因子. 其中, 性别可用 0 和 1 表示; 年龄可以做分段量化处理; 买药频率、花费和数量可以原始数据统计得到; 科室字段可用科室 ID 代替; 因同一个病人可以对应多个医嘱项, 故医嘱项不能直接作为输入因子, 可用医嘱重复因子替代表示, 计算公式为: 医嘱重复因子=医嘱项总数/医嘱种类数, 当该因子越小时, 表示该病人的相对重复率也高, 涉嫌医保欺诈的可能性也越高.

经过数据预处理, 用统计学的方法建立识别模型. 对于欺诈数据和非欺诈数据, 可以分别赋予期望 1 和 0. 以此将总体样本划分为欺诈样本数据 (Y 类) 和正常样本数据 (N 类).

3.2 BP 神经网络的构建

对于一般的模式识别问题, 三层网络即可很好的达成识别问题任务. 本研究选取一层隐含层, 即采用三层神经网络结构进行欺诈识别. 隐层节点根据经验公式: $n_2=2\times n_1+1$(n_i 是第 i 层网络节点数), 确定数量为 19 个[15].

最后的网络输出层是由是否欺诈 (0 和 1) 组成的一维矩阵, 0 代表没有欺诈, 1 代表欺诈. 值越高, 表明该记录欺诈可能性越高. 最终建立的神经网络如图 1 所示.

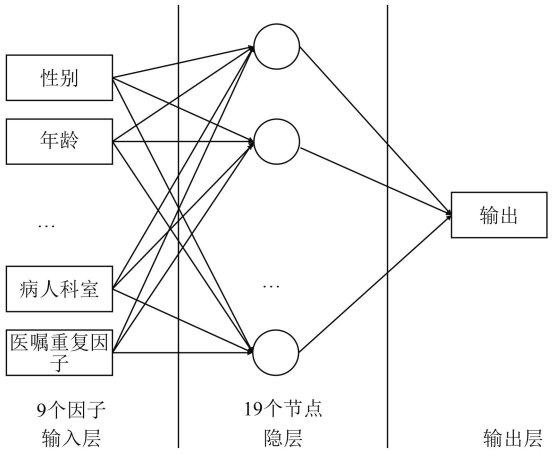

图 1 BP 神经网络结构图

3.3 BP 神经网络模型测试

从初步筛选阶段所划分的欺诈类和疑似欺诈类中分别随机抽取 2000 个样本组合作为 BP 神经网络的训练样本, 并分别赋予期望值 0 和 1. BP 神经网络的运算采用 matlab 软件. 在 BP 神经网络的建立中, 隐含层采用正切函数“tansig”作为激活函数, 输出层采用对数函数“logsig”作为激活函数, 以保证最后的输出结果值域在 0 到 1 范围内. 考虑到 LM 算法在稳定性等方面表现优良, 本文采用 LM 算法的 trainlm 函数对网络进行训练. 设定最大循环次数 5000 次, 目标误差 0.001, 最小梯度 1e–20.

对训练样本数据建立的神经网络如图 2 所示.

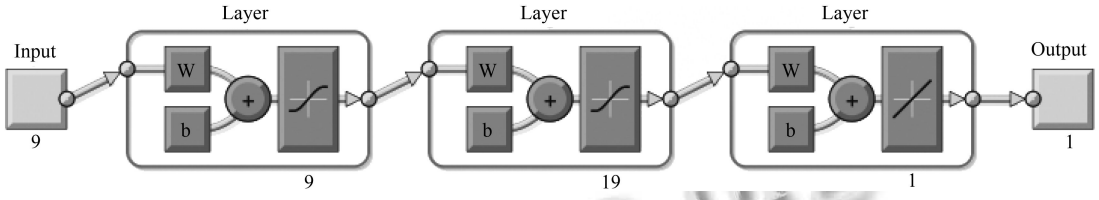

图 2 Matlab 构建的 BP 神经网络图

样本的训练经过 86 次迭代, 用未加入噪声的训练样本检测 BP 神经网络, 将输出的结果与未加入噪声的训练样本进行匹配. 经程序运行检验, 训练样本数据的识别准确率达 93%.

通过图 3 可以直观的发现, 期望输出的‘o’与检测时的实际输出‘*’基本吻合, 而只有极少数的点出现了误判, 从而说明该神经网络是可靠的.

3.4 模型的改进

本研究初步选用了 9 个欺诈因子作为模型的输入(如表 1 所示), 但是其中各因子对结果的影响程度是不用的, 通过引入 Logistic 回归分析剔除一些影响较弱的因子. 若用 p 来表示欺诈事件发生的概率, 同时将 p 看做自变量 x 的函数, 则因子变量和自变量识别因子的关系保持传统回归模式.

$$logitp=\beta_0+\beta_1x_1+\beta_2x_2+\beta_2x_2+\cdots+\beta_kx_k$$
$$=\beta_0+X\beta$$

其中, $x_1,x_2,x_3,\cdots,x_k$分别为 k 个识别因子, 为β_0常数项, $\beta_1,\beta_2,\beta_3,\cdots,\beta_k$分别为 k 个自变量的回归系数.

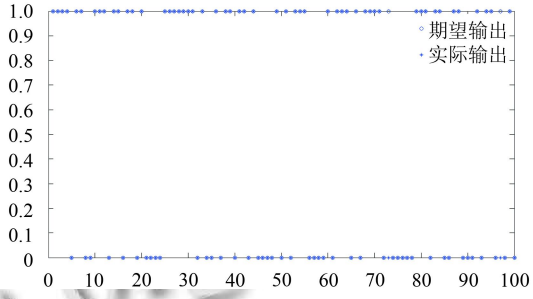

图 3 BP 神经网络欺诈预测结果

表 1 9 个欺诈因子

变量	回归系数	标准误差	Wald 卡方	Sig.	Exp(B)
性别	–0.146	0.135	1.170	0.279	0.865
年龄	0.14	0.004	13.722	0.000	1.014
买药频率	–0.144	0.010	198.343	0.000	0.866
买药总花费	–0.009	0.000	828.017	0.000	0.991
买药总数	0.000	0.000	3.550	0.060	1.000
下嘱科室	6.243	1387.180	0.000	0.996	514.360
执行科室	–0.002	0.001	2.570	0.109	0.998
病人科室	–6.246	1387.180	0.000	0.996	0.002
医嘱重复因子	–1.914	0.440	18.902	0.000	0.147

本文用 SPSS Statistics 22 对输入因子进行 Logistic 回归分析, 提取其中的显著因子, 分析结果如表 2.

分析可得, 总的预测准确率为 94.1%, 其中, 买药总花费和买药频率作为主要影响因子, 医嘱重复因子、年龄、买药总数是次要影响因子, 而下嘱科室、病人科室、执行科室以及性别几乎对模型无影响, 可以作为无效因子剔除, 从而提高输出结果的准确度.

表 2 Logistic 回归分析结果

已观测		已预测		
		是否欺诈		百分比校正
		否	是	(%)
是否欺诈	否	1899	101	95.0
	是	136	1864	93.2
总计百分比 (%)				94.1

此外, 上文研究是用欺诈数据 Y 和正常数据 N 共同训练神经网络. 但是考虑到样本数据的稀缺, 以及实际情况下, 医保欺诈数据在海量数据中只占很小的比重, 因此本文采用一个新的角度, 只用正常数据 N 来训练神经网络.

BP 神经网络有一个很重要的特性: BP 神经网络能够模拟任何连续曲线. 因此, 如果只用正常数据来训练神经网络, 模型预期将稳定的输出 0. 模型训练好后, 用待测的数据输入, 神经网络会输出趋近于 0 的数据. 如果输入的是符合常规情况的数据, 则输出的结果与期望值 0 的误差会小于神经网络学习时的误差, 若输入的是疑似欺诈的数据, 则输出结果与期望值 0 的误差会成倍增加. 这种方式可以很好的解决大部分数据不存在欺诈的问题, 同时也降低了复杂度.

用正常数据作为训练样本、经过回归分析剔除弱因子后剩余的 5 个因子作为输入因子按照上文步骤重新训练网络, 设定目标误差为 0.001. 表 3 是将测试数据输入训练后的网络模型得到的计算结果.

表 3 测试数据计算结果表

ID	年龄	购买频率	购买花费	购买药品总数	医嘱重复因子	期望输出	实际输出误差
667712	31	2	417.39	35	1	0	–0.000 911 625 1
691223	44	3	603	271	1	0	0.000 428 585 3
236247	56	1	73	288	1	0	–0.000 400 068 2
209394	8	11	345.92	48	0.7272	0	–0.000 298 218 8
565325	38	4	185.8	144	1	0	0.000 378 815 5
670090	39	2	191.65	40	1	0	–0.000 446 708 9
655501	46	2	40.8	160	1	0	–0.000 961 842 1
352740	39	3	70.01	2	1	0	0.0003306847
253694	34	2	60.04	60	1	0	–0.000 351 639 5
679835	51	5	243.62	138	0.6	0	–0.000 247 256 8
680538	26	14	182.06	210	1	0	**0.001 503 257 1**
271611	49	1	312.94	16	1	0	–0.000 400 068 2
547946	3	1	49.28	1	1	0	–0.000 247 256 8
688250	32	1	59	10	1	0	0.000 430 373 5
694091	36	2	94.78	36	1	0	–0.000 450 359 7
532467	10	2	27	40	1	0	–0.000 248 523 4
518426	29	2	42.66	6	0.5	0	0.000 478 504 3
689074	21	4	0	0	1	0	–0.000 299 485 4
614778	29	2	99.29	98	1	0	–0.000 197 561 5
619779	41	1	59.76	155	1	0	0.000 478 504 3
590998	12	4	20	2	1	0	–0.000 500 576 6
659306	5	3	114.61	43	1	0	–0.000 396 342 9
299284	51	5	96.22	88	0.8	0	0.000 377 027 3
512071	79	2	163.45	37	1	0	–0.000 251 056 6
655197	26	1	14.02	200	1	0	–0.000 982 133 4
227205	67	5	2.33	31	1	0	0.000 428 585 3
484621	38	2	1215.7	2193	1	0	**–0.001 140 357 9**
304660	18	1	116.36	1	1	0	–0.000 346 647 6
190777	49	11	965.77	1064	0.8181	0	**0.001 102 266 8**

由表 3 可见, 大部分的数据的实际输出误差小于 0.001, 表明该神经网络能够较好的模拟正常情况下的医保消费. 其中, ID 为 680538、484621、190777 的病人的误差绝对值超过了 0.001, 其中 ID 为 680538 的病人误差绝对值最大, 对这几项数据进行进一步审核, 发现符合医保欺诈特征, 可见该神经网络对一些明显的医保欺诈具有较好的识别性.

只用正常数据训练网络的方法对于医保欺诈的识别有重要参考价值. 尤其是在欺诈识别应用初建, 对欺诈数据样本积累不足的情况下. BP 神经网络运算结果

受数据样本选取的影响较大, 若训练样本选取不当, 可能很难达到预期效果. 因此, 若欺诈样本积累不足, 用传统方法训练网络最终的识别效果可能不佳. 本文方法适合在医保欺诈识别工作开展的初期, 加快对欺诈数据的筛选工作, 可以大幅降低人工审核筛选样本的工作. 待欺诈样本累计一定程度后, 可以考虑选择用常规的欺诈和非欺诈样本训练网络或其他更优方法进行欺诈识别.

4 结语

本文对比众多保险欺诈识别的研究, 采用 Logistic 回归和 BP 神经网络相结合的方式构建模型, 并只用正常数据训练神经网络模拟函数曲线的方式来进行医疗保险欺诈的识别. 该方法剔除了对欺诈识别弱影响的因子, 解决了欺诈数据稀缺对模型训练的影响. 实证表明, 该模型具有较高的准确性, 应用该方法进行医保欺诈识别是可行的.

但研究中还存在一些不足. 模型的影响因子还不够健全, 从而影响模型的识别效果. 此外数据来源的有限性也导致模型泛化能力欠佳. 对以上方面的改进将能进一步提升 BP 神经网络对医疗保险欺诈的识别效果.

参考文献

1 Artís M, AyusoM, Guillén M. Detection of automobile insurance fraud with discrete choice models and misclassified claims. The Journal of Risk and Insurance, 2002, 69(3): 325–340. [doi: 10.1111/1539-6975.00022]

2 Brockett PL, Derrig RA, Golden LL, *et al*. Fraud classification using principal component analysis of RIDITs. The Journal of Risk and Insurance, 2002, 69(3): 341–371. [doi: 10.1111/1539-6975.00027]

3 Pinquet J, Ayuso M, Guillén M. Selection bias and auditing policies for insurance claims. The Journal of Risk and Insurance, 2007, 74(2): 425–440. [doi: 10.1111/jori.2007.74.issue-2]

4 叶明华. 构建我国机动车保险欺诈识别的指标体系——基于江、浙、沪机动车保险索赔样本数据. 保险研究, 2010, (4): 83–87.

5 叶明华. 基于 BP 神经网络的保险欺诈识别研究——以中国机动车保险索赔为例. 保险研究, 2011, (3): 79–86.

6 Viveros MS, Nearhos JP, Rothman MJ. Applying data mining techniques to a health insurance information system. International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. 1996. 286–294.

7 Chiu CC, Tsai CY. A web services-based collaborative scheme for credit card fraud detection. IEEE International Conference on E-Technology, E-Commerce and E-Service. IEEE, 2004. 177–181.

8 何俊华, 张静谊, 熊赟, 等. 医保就医聚集行为挖掘. 计算机应用与软件, 2011, 28(7): 79–81.

9 陈亚琳, 王旭明. 基于数据挖掘的医保欺诈预警模型研究. 电脑知识与技术, 2016, 12(11): 1–4.

10 唐璟宜, 孙有坤, 周海林. 医保欺诈行为的主动发现. 合作经济与科技, 2016, (16): 188–190. [doi: 10.3969/j.issn.1672-190X.2016.16.091]

11 Maier HR, Dandy GC, Burch MD. Use of artificial neural networks for modelling cyanobacteria Anabaena, spp. in the River Murray, South Australia. Ecological Modelling, 1998, 105(2): 257–272.

12 林源. 国内外医疗保险欺诈研究现状分析. 保险研究, 2010, (12): 115–122.

13 Liou FM, Tang YC, Chen JY. Detecting hospital fraud and claim abuse through diabetic outpatient services. Health Care Manag Sci, 2008, 11(4): 353–358.

14 Maes S, Tuyls K, Vanschoenwinkel B, *et al*. Credit card fraud detection using Bayesian and neural networks. In: Maciunas RJ, ed. Interactive Image-guided Neurosurgery. American Association Neurological Surgeons. 1993. 261–270

15 CSDN. BP 神经网络算法学习——基础理论 1. http://blog.csdn.net/fuwenyan/article/details/53893817. [2016-12-27].