文章编号:1671-4598(2018)04-0167-04 DOI:10.16526/j.cnki.11—4762/tp.2018.04.044 中图分类号:TP393.4 文献标识码:A

失稳网络医保信息欺诈检测算法研究

吴 剑
(北京交通大学 经济管理学院,北京 100044)

摘要:传统医保信息欺诈检测算法存在运行时间长、效率低的问题,无法保障患者医保信息安全,为了解决该问题,采用基于随机森林算法对失稳网络医保信息欺诈行为进行检测;通过混合抽样可抽取在失稳情况下的数据,并建立非平衡数据分类算法抽样机制;进行迭代随机森林数据计算,采用多数投票法构建基分类器,并以此为基础筛选异常数据;利用模型实现该算法对医保信息欺诈检测;设计对比实验,验证该算法有效性;通过实验结果可知,基于随机森林算法运行时间较短、效率高。

关键词:失稳网络;医保信息;欺诈;随机森林算法;混合抽样;基分类器

Research on Information Fraud Detection Algorithm for Unstable Network Medical Insurance

Wu Jian
(School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Traditional health insurance information fraud detection algorithm has many problems such as long running time and low efficiency, which cannot guarantee the safety of medical insurance information. In order to solve this problem, we use random forest algorithm to detect medical fraud information in unstable network. Through extracting mixed sampling instability in the case of data, and the establishment of the imbalanced data classification algorithm for iterative sampling mechanism; random forest data, builds a classifier using voting method, and on the basis of screening of abnormal data; the algorithm of insurance fraud detection using information model. Design comparison experiment to verify the effectiveness of the algorithm. The experimental results show that the run time based on random forest algorithm is shorter and more efficient.

Keywords: unstable network; medical insurance information; fraud; random forest algorithm; mixed sampling; base classifier

0 引言

目前我国老龄化程度日益增加,对于医疗保险需求也越来越多,设立完善医疗保险机构是保证人们生活水平向更好方向发展基础,也是维护社会环境强大后盾。医疗保险不但能够改善人们生活水平,还会促进我国经济向稳定、可持续方向发展。该结构建立目的是保障人们受到基本医疗保障需求,其包含了与人们日常生活相关养老、失业、医疗等范围,通过医疗机构向个人或单位筹集专款,如果出现资金不足问题,政府还可提供专项资金来支撑。近几年,医疗保险行业快速发展,促使每3个人就会有1人具有医疗保险,同时,医保欺诈问题逐渐增多。由于欺诈行为多种多样,尽管使用欺诈识别技术,也很难保障每个人的医疗保险都不会受到欺诈行为,无疑是对我国经济造成较大影响。犯罪人员通过保险欺诈行为,违反相关法律法规,获取医保基金,由于我国目前有关医保安全整治工作发展较迟,导致能够安全使用医保人群较少,因此常常出现医保欺诈行为[1-2]。医保欺诈具有多种表现方式,其行为涉及到多种机构或个体,由于欺诈行为会造成治疗费用增加,为此,需对失稳网络医保信息欺诈行为进行检测。采用传统检测算法存在运行时间长、效率低问题,无法保障患者医疗保险使用安全。

针对传统算法存在的问题,提出了基于随机森林算法对失稳网络医保信息欺诈行为进行检测。选择分布点种类和实际应用情况进行分析,借鉴专家领域对分布点进行深入细致研究,通过实验对比结果可知,采用该算法可快速对违规记录进行识别,并且检测效率较高。

收稿日期:2018-02-09; **修回日期**:2018-02-27。

作者简介:吴 剑(1994-),男,陕西安康人,主要从事企业信息化与医疗信息化方向的研究。

1 基于随机森林医保信息欺诈检测算法

结合医保数据在网络不稳定条件下不平衡属性,采用随机森林算法对医保信息欺诈行为进行检测。通过混合抽样可保证非平衡数据在失稳情况下平衡化处理,使用森林分类方式,经过迭代运算对数据进行平衡化处理,通过分类性能对基分类器进行选择,进而提高算法检测准确性,利用模型实现该算法对医保信息欺诈检测[3]。

1.1 数据处理方式

1.1.1 随机森林抽取数据

多个决策簇构成了随机森林集成算法,在每个簇进行训练完成森林集合之前,需使用前端框架有放回的对数据进行抽样,从原始样本数据中抽取数量相同样本作为训练集合,经过抽样获取的数据组成里包(inBag)样本集合,将没有抽取的样本数据组成外包(outBag)样本集合,其中里包(inBag)样本集合为基分类器的训练子集[4]。随机森林抽取数据获取的训练集合流程如图1所示。

每次迭代运算训练子集的数量与原始样本数量一致,但是采用有放回的抽样方式会出现数据被重复抽到的问题,为此,当训练数据为非平衡数据时,其数量差异性会导致原始样本数据集合抽样时,少量样本一次都没有抽到。如果使用较少样本训练子集作为基分类集合,将导致基分类器无法对样本进行识

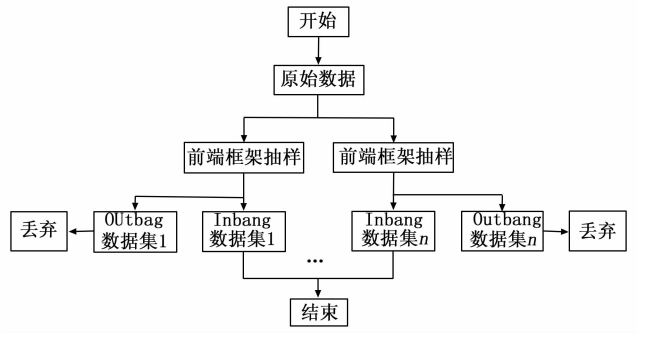

图 1 随机森林获得训练子集流程图

别，为此需构建非平衡数据分类算法抽样机制。

1.1.2 非平衡数据分类算法抽样机制建立

通过对非平衡数据采用 smote 方式，即通过合成新的少数类样本，对每个少数类样本 a，从它最近邻中随机选一个样本 b，然后在 a、b 之间连线上随机选一个点作为新合成的少数类样本。通过该方式增加样本数量，可以有效保障数据拟合程度，为此在每次迭代运算过程中，使用 smote 方式可对全部数据进行重构，进而解决样本中不平衡问题。由于采用 smote 方式生成的新样本存在随机性，对于增加样本差异性，需使用分类器进行属性互补[5]。经过处理之后的随机森林训练样本子集算法具体流程如下所示：

设样本输入的初始集合为 F，进行随机森林迭代运算的次数为 S；输出的训练样本子集为 inBag、测试样本子集为 outBag，具体机制为：

For $t=1$ to T;

1）通过对非平衡数据处理，采用 smote 方式对初始数据样本集合进行平衡化处理：

$$F = \text{smote}(F) \tag{1}$$

2）经过步骤 1）中获取的相对平衡数据样本集合，采用有放回抽样方法获取输出训练样本子集；

$$inBag = \text{bootstrap}(F) \tag{2}$$

3）将未抽取到的样本子集作为测试样本子集。

$$\text{outBag} = F - \text{inBag} \tag{3}$$

1.2 基分类器的选择

1.2.1 随机森林组合基分类器特点

进行迭代随机森林数据计算过程中，需按照相同方式来抽取子集，基分类器按照并行或独立方式生成彼此之间联系性，由于基分类数据地位是相同的，因此采用多数投票法构建基分类器，如图 2 所示。

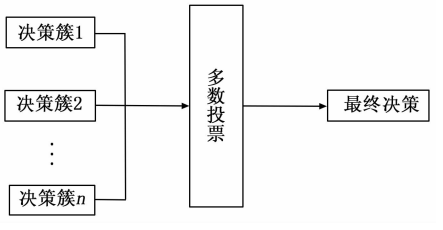

图 2 随机森林组合基分类器示意图

由图 2 可知：在随机森林组合中对分类器进行集成，不需考虑分类器自身属性，如果分类器中的分类属性较多时，需将集成效果降低，只有基分类器中拥有较低集成效果时，才可提高算法准确率。为此，待集成决策簇数量越多并不代表随机森林对数据分类效果就越好[6-8]。

1.2.2 按分类性能筛选基分类器

根据数据随机组合，基分类器在集成过程中，会将性能较差基分类器去除，进而降低集成效果，为此在对分类器进行筛选时，需将性能较差分类器剔除，只使用效果较好集成分类器。由于在迭代运算过程中，抽取到的里包（inBag）样本集合作为训练子集，采用 smote 方式进行反复样品，大约有 40%的数据样本没有被抽取到，外包（outBag）样本集合并没有完全参与簇的决策之中，因此在随机森林进行迭代运算时，需使用抽取到的里包（inBag）样本作为基本训练的子集，然后将没有被抽取到的外包（outBag）样本集合作为测试集合[9]。

针对性能较好的基分类器进行数据分类时，需采用异常数据样本统计量值作为对非平衡数据分类的评价标准，综合统计量中的准确率和召回率，评价每一个决策簇的特点，统计量值越高，代表分类效果就越好。按分类性能筛选的基分类器具体步骤如图 3 所示。

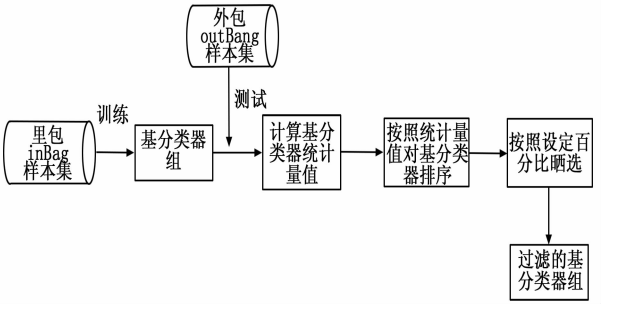

图 3 按分类性能筛选基分类示意图

按照分类性能筛选出性能较好的基分类器，并由上述图 3 可知，在随机森林算法进行迭代运算过程中，分别使用抽取到的里包（inBag）样本作为基本训练子集，将没有被抽取到的外包（outBag）样本集合作为测试子集对该基分类器进行测试。经过迭代运算后，所有的基分类器都可按照统计量方式进行倒序排序，并将分类效果较好分类器放置在最前方，根据设定函数以降序形式返回表中最前面的几行，这些行的累积合计至少要达到指定百分比，并筛选其中一部分效果较差基分类器。如果将 100 个基分类器同时进行倒序排列，那么函数百分比为 80%，则有 80 个基分类器被选择，剩余的 20 个基分类器被剔除，进而提高集成分类效果。

1.3 算法模型

根据上述基分类器选择以及对数据处理，可构建随机森林算法模型，在进行迭代运算过程中，需通过既定原始数据直接使用反复抽样方式进行数据抽取，并将抽取到的数据作为训练子集，完成相应簇的决策。迭代运算结束后，将所有决策树都参与到集合之中，由于不同决策树是相互独立的，且不具有任何属性，因此按照投票结果完成簇的决策集合[10]。即使在不同决策簇中的分裂中心节点处具有特征集合的筛选功能，但是由于数据未被进行特征化同化，所以不同将该部分直接展示出来，基于此，构建随机森林算法模型，如图 4 所示。

针对医保信息需从数据层次上和算法层次上分别对不平

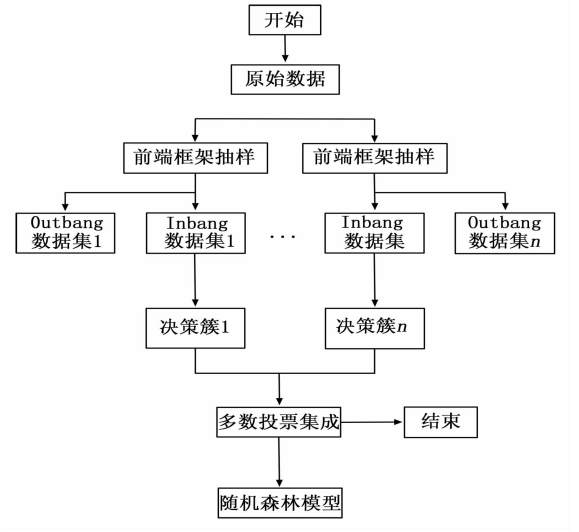

图 4　随机森林算法流程图

衡数据进行平衡化处理，并利用集成思路完成欺诈行为的检测。其中，在数据层次上，由于网络不稳定，导致医保信息中也存在着一种不平衡的数据，为此需对数据进行平衡化处理，利用 smote 方式进行混合抽样来改善数据不平衡所造成的分类效果差的问题，进而增加异常数据，为迭代运算 smote 方式所产生新数据进行分类处理，并通过前端框架进行有放回差异性训练处理，进而提高数据差异性，便于检测；在算法层次上，对数据集成之前，需进行两次筛选，一次是对决策簇数据进行筛选，一次是对统计量值进行筛选。由于训练样本存在差异性，促使决策簇形成也是不同的，每次先对决策簇数据进行筛选，将差异性较大数据识别出来，并剔除。根据基分类性能统计量衡量指标，获取到每一个基分类器统计量值后，按照依次排列，筛选出性能较好分类器，并进一步处理。通过不一致度量值作为衡量决策簇差异性重要指标，需综合考虑分类器相似度，如果相似度低于正常阈值时，再考虑决策簇分类性能，提出多余模型，可提高分类性能，又保证算法检测效果。

2　实验

采用基于随机森林算法对失稳网络医保信息欺诈行为进行检测，为了验证该检测算法的有效性，将传统算法与该算法进行对比验证，以此提高该算法的可靠性，具体实验内容与结果如下所示。

2.1　实验过程

在医保数据中存在欺诈行为的主要原因就是参与人信息与正常法规数据相比出现误差，导致在医保数据中出现分布的现象。根据某次参与医保的患者数据，在较短时间内有频繁住院的情况，或者存在医保欺诈事件，采用随机森林检测算法对医保欺诈事件进行有效识别，进而达到实验的目的。实验选取的数据来源于某市农村信用社与医疗机构合作的办公室所提供的 2017 年 1 月 1 日至 2017 年 12 月 31 日的 11215 条住院记录，通过人工审计可确定其中 30 条记录为异常信息，以此作为两种检测方法准确评估标准。在内存大小为 8G 的计算机上使用 C++语言实现两种算法的检测。

2.1.1　第一阶段

首选从上述 11215 条住院记录中随机选取一条诊断记录作为聚类簇的中心，并将其赋予阈值，经过计算和比较，可最终集合划分成 8 个聚簇，分别为 1138、3619、1123、50、672、1668、1123 和 1822 条住院记录，计算 8 个聚簇相关的 3 个属性，分别是：簇基数、簇半径、簇分布点可能性指标。对聚类结果进行统计与排列，结果如表 1 所示。

表 1　聚簇相关属性表

编号	簇序号	簇基数	簇半径	指标
1	A3	1138	1725.362	1.903
2	A1	3619	1125.513	1.023
3	A7	1123	1692.245	0.934
4	A5	50	728.203	0.813
5	A4	672	35.861	0.775

2.1.2　第二阶段

在实验进行过程中，参与的变量因素包络：住院总费用、实际补偿费用、自付费用、药品费用、可实际补偿医药费用、医药使用率、住院时间等，还包括经过分布离散化处理之后的住院登记、患者种类、患者家庭种类、医保补偿方式等因素。在该阶段可对已经编号的 5 个聚簇进行逐一扫描，记录每个分布的因子，构建优先级队列，实现邻近分布点的快速搜索与排列。为了保证患者信息安全，将历史记录中的患者姓名进行了编号处理，并截取部分历史记录数据来显示具体分布点检测结果，如表 2 所示。

表 2　具体分布点检测结果

姓名	疾病名称	住院总费用	实际补偿费用	自付费用	患者家庭种类	患者种类	医院级别	住院时间/天
1	肝癌	166133.640	55140.050	110993.590	农户	正常	省级以上机构	27
2	脑梗	1004.650	802.110	140.452	贫困户	低保	乡镇级机构	10
3	先天性二尖瓣闭锁不全	124916.200	42911.450	80050.420	农户	正常	省级以上机构	23
4	冠动脉硬化心脏病	188125.610	58917.550	129246.960	农户	正常	省级以上机构	33
5	冠心病	692.300	513.780	160.103	贫困户	低保	乡镇级机构	7
6	脑梗	780.200	600.180	152.813	农户	正常	乡镇级机构	20
7	脑梗塞	2159.150	2094.210	323.910	农户	正常	乡镇级机构	20
8	糖尿病	141901.610	37301.920	104287.560	农户	正常	地市级机构	29
9	脑梗死后遗症	1124.099	921.650	201.310	五保户	三结户	乡镇级机构	10
10	盆腔炎	100.560	93.510	100.021	农户	正常	乡镇级机构	8

表 2 中共有异常信息 20 条，查准率为 80%，该准确率可用于比较传统检测方法与本文研究检测方法的可靠性。

2.2 实验结果与分析

2.2.1 检测运行时间对比结果与分析

在医保数据集上，将传统检测方法与本文检测方法对医保欺诈行为检测所运行时间进行对比，结果如图 5 所示。

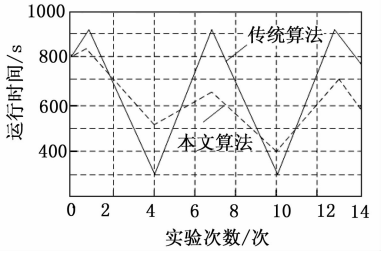

图 5 两种检测算法运行时间对比结果

由图 5 可知：当实验次数为 2 次时，两种算法检测运行时间一致，随着实验次数增加，传统算法运行时间不稳定，上下波动幅度较大，虽然在实验期间有几次运行时间比本文算法时间要少，但网络不稳定，导致传统算法最后运行时间稳定在 760 s 左右；而本文算法运行时间相对稳定，上下波动幅度并不大，最终本文算法运行时间稳定在 580 s 左右。由此可知，采用基于随机森林算法对失稳网络医保信息欺诈行为进行检测运行时间较短。

2.2.2 算法检测效率对比结果与分析

根据上述实验过程可知，表 2 中共有异常信息 20 条，查准率为 80%，为了使结果更具有可靠性，将传统算法与本文算法检测效率进行对比，从算法检测效率和分布点检测数量两个方面进行比较，并采用 C++语言在统一平台下编译，对比结果如图 6 所示。

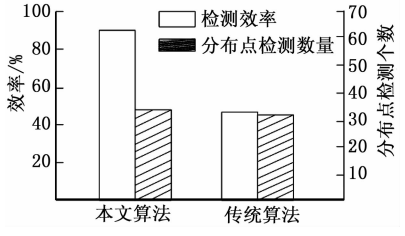

图 6 两种检测算法检测效率对比结果

由图 6 可知：采用本文算法检测效率高于传统算法 40%，且分布点检测个数比传统要多，综合来看，本文算法具有较好实用性。

2.3 实验结论

根据上述实验内容，可得出结论，由于医保欺诈记录包含多个相关属性，采用传统算法不能全面反映实际发生情况，更无法直接筛选出分布点，采用基于随机森林算法对失稳网络医保信息欺诈行为进行检测，可快速对违规记录进行识别，辅助管理人员完成审核工作，通过对比结果可知，采用本文检测算法运行时间较短、效率较高，且分布点检测个数比传统要多，综合来看，本文算法具有较好实用性。

3 结束语

针对数据挖掘流程，为设计医保欺诈检测算法奠定理论基础，选择分布点种类和实际应用情况进行分析，并对分布点进行深入细致研究，比较传统检测算法可知，存在运行时间长、效率低的问题，明确不同方法处理数据集合的使用范围，掌握分布点检测思想。借鉴专家领域，选择具有数据属性参与计算方式，通过实验结果可知，采用基于随机森林算法对失稳网络医保信息欺诈行为进行检测，可快速对违规记录进行识别，存在运行时间较短、效率较高，且分布点检测个数比传统要多等优势，检测效果较好。

参考文献：

[1] 孙 菊，甘银艳. 合作治理视角下的医疗保险反欺诈机制：国际经验与启示 [J]. 中国卫生政策研究，2017，10 (10)：28－34.

[2] 梅丽萍. “聪明监管”：基本医疗保险监管的模式和路径选择 [J]. 中国卫生经济，2016，35 (6)：13－18.

[3] 李亚子，虞昌亮，吴春艳，等. 新型农村合作医疗与城镇居民基本医疗保险制度整合中信息系统整合技术路线研究 [J]. 中国卫生经济，2017，36 (1)：34－36.

[4] 侯 俐，焦 锏. 经济信息欺诈与经济信息政策分析——评《网络经济时代的信息政策》[J]. 宏观经济管理，2017，25 (1)：24－26.

[5] 彭 玲，阳作松，杨新艳，等. 基于信息技术的医院医保闭环式管理 [J]. 中国医院管理，2017，37 (1)：59－61.

[6] 王雄军，张冰子. 我国医保改革的地方经验评述与启示 [J]. 中国党政干部论坛，2016，26 (5)：58－62.

[7] 王娟丽，邓明文. 西藏城乡居民基本医疗保险制度并轨问题探讨 [J]. 中国卫生经济，2017，36 (7)：31－34.

[8] 郑先平，傅强辉，刘 雅. “互联网＋”背景下医疗保险异地结算路径优化 [J]. 卫生经济研究，2017，15 (5)：63－65.

[9] 艾丽唤，吴荣海，肖 黎，等. 基于风险调整的基本医疗保险门诊统筹按人头付费标准测算研究 ——以深圳市为例 [J]. 中国卫生政策研究，2017，10 (9)：12－14.

[10] 陈 颖，魏永祥，刘海燕，等. 商业医疗保险在公立医院中的实践 [J]. 中华医院管理杂志，2016，32 (2)：102－104.

(上接第 166 页)

参考文献：

[1] 谭 民，王 硕. 机器人技术研究进展 [J]. 自动化学报，2013，39 (7)：963－972.

[2] 张岩岩，侯媛彬，李 晨. 基于人工免疫改进的搬运机器人蚁群路径规划 [J]. 计算机测量与控制，2015，23 (12)：4124－4127.

[3] 吕太之，赵春霞，夏平平. 基于同步可视图构造和 A＊算法的全局路径规划 [J]. 南京理工大学学报 (自然科学版)，2017，41 (3)：313－321.

[4] 孙 炜，吕云峰，唐宏伟，等. 基于一种改进 A* 算法的移动机器人路径规划 [J]. 湖南大学学报 (自科版)，2017，44 (4)：94－101.

[5] 陶重犇，刘壮宇，孙云飞. 基于嵌入式系统的搬运机器人设计与路径规划研究 [J]，计算机测量与控制，2016，24 (8)：215－217

[6] 宗成星，陆 亮，雷新宇，等. 一种基 A* 算法的空间多自由度机械臂路径规划方法 [J]. 合肥工业大学学报：自然科学版，2017，40 (2)：164－168.

[7] 裴振兵，陈雪波. 改进蚁群算法及其在机器人避障中的应用 [J]. 智能系统学报，2015 (1)：90－96.

[8] 黄 辰，费继友，刘 洋，等. 基于动态反馈 A* 蚁群算法的平滑路径规划方法 [J]. 农业机械学报，2017，48 (4)：34－40.

[9] 温素芳，郭光耀. 基于改进人工势场法的移动机器人路径规划 [J]. 计算机工程与设计，2015 (10)：2818－2822.

[10] 石为人，黄兴华，周 伟. 基于改进人工势场法的移动机器人路径规划 [J]. 计算机应用，2010，30 (8)：2021－2023.