基于随机森林的保险欺诈识别研究

曹诗琦
（南开大学金融学院，天津 300350）

摘要：随着保险业的不断发展，保险欺诈情况也变得越来越严重，保险欺诈严重违背了保险的最大诚信原则，不仅给保险公司的利益带来了损害，还扭曲了保险定价的机制，损害了广大诚实消费者的利益，同时也给社会带来了损失。保险欺诈识别能帮助保险公司在承保，理赔时采取一定的措施，如对投保方进行一定约束，对索赔进行特别关注或进一步调查等，能够有效减少保险欺诈。文章首先从经济、法律、保险公司和社会这四个不同的方面分析了保险欺诈产生的原因，然后介绍了随机森林方法，再构建模型，用随机森林产生差异度，根据差异度进行聚类，将数据分类，并将得到的结果与用欧氏距离进行聚类的分类结果进行比较，研究随机森林方法用于保险欺诈识别的可行性。

关键词：保险欺诈；欺诈识别；随机森林；聚类
中图分类号：TP391　　**文献标识码：**A　　**文章编号：**2096-9759(2023)06-0017-03

Research on insurance fraud identification based on random forest

CAO Shiqi
(School of Finance of Nankai University, Tianjin 300350, China)

Abstract: With the development of the insurance industry, the situation of insurance fraud is becoming more and more serious. Insurance fraud violates the principle of utmost good faith, and it not only damages the interests of insurance companies, but also distorts the insurance pricing mechanism and harms the interests of honest consumers, and also brings losses to the society. Insurance fraud identification can help insurance companies to take certain measures when underwriting and settling claims, such as certain restrictions on the insured, special attention to claims or further investigation. It can effectively reduce insurance fraud. This thesis firstly analyzes the causes of insurance fraud from four aspects: economic reasons, legal reasons, insurance companies' reasons and social reasons. Then, the random forest method is introduced in this thesis, and then the model is built. The difference degree is generated by the random forest, and then the difference degree is clustered, so as to classify the data, and the results of clustering are compared with Euclidean distance to study the feasibility of the random forest method in the identification of insurance fraud.

Keywords: Insurance fraud; Fraud identification; Random forest; Clustering

1 研究背景及意义

近年来，我国保险业总体来说保持了平稳增长的态势，增速趋缓。截至 2022 年 6 月 6 日，我国共有保险机构 350 家。2021 年全国原保险保费收入 44900.17 亿元，赔付支出 15608.64 亿元，与 2020 年相比增长 12%[1]。

在保险业发展的过程中，由于信息的不对称，保险欺诈情况变得越来越严重。根据国际保险监管者协会（IAIS）测算，全球每年约有 20%~30%的保险赔款涉嫌欺诈。中国人民大学教授孟生旺曾指出，我国保险公司每年赔款中至少有 10%~20％属于保险欺诈。

保险欺诈给消费者、企业、政府和社会都带来了非常严重的损失。保险欺诈增大保险公司赔付支出，损害保险公司利益，同时，赔付率的上升导致保险产品价格上升，扭曲定价机制，欺诈索赔和打击欺诈增加了保险公司的成本，保险公司又将这部分增加的成本向投保人转移，从而损害了众多诚实消费者的利益。其中，受害的企业又通过提高商品和服务的价格，将保费上涨的成本转嫁给客户，从而使许多无辜者的利益受到损害。保险欺诈严重违背了保险经营的最大诚信原则，动摇了保险业的基础，使保险市场的秩序被打乱，有损保险业的形象。另外保险欺诈的情形正在变得越来越复杂，受到影响的往往不只是保险业，而是多个行业。

本文将随机森林应用于保险欺诈识别研究，通过随机森林对保险数据进行分类，验证随机森林在保险欺诈识别中的可行性，为保险行业欺诈识别研究做出一定贡献。

2 保险欺诈成因分析

2.1 经济原因

2.1.1 损失程度度量的主观性

事故发生后，对损失程度的度量要做到公正客观和准确通常有一定的难度。如在火灾事故中，被烧毁的财产数量、烧毁的程度、剩余残值的多少等等，界定起来有一定困难，从而使保险欺诈的发生有了可能性。

2.1.2 低风险，高回报特点

与其他的犯罪活动相比来说，保险欺诈常常被公众认为是低风险、高回报的。因为保费与保险赔款的差距，有些人认为保险欺诈能获得高额的回报，并且认为即便欺诈不成功，结果要么拒赔，要么解除保险合同，不会付出太大的代价，而并没有意识到保险欺诈行为的严重性。

2.2 法律原因

法律上对于保险欺诈还不够重视，对欺诈行为的打击程度不够。美国通常更重视打击毒品、暴力和其他受到更多关注的犯罪行为。而我国还没有法律规定保险公司拥有的调查权，因此在很多案件中保险人没有办法取证，不利于欺诈行为

收稿日期：2023-03-04
作者简介：曹诗琦（1998-），女，四川成都人，研究生，硕士，主要研究方向：机器学习与精算模型。

的识别，给保险欺诈的产生提供了可乘之机。同时，我国对欺诈行为的处理侧重于解除保险合同，通常对于欺诈未遂的行为人，因为其行为尚未构成犯罪，只进行批评教育，并不追究刑事责任。法律约束一定程度上的不足构成了保险欺诈产生的原因。

2.3 保险公司原因

2.3.1 保险标的与保险人的分离

保险标的大部分时间完全处在投保方的掌控之中，保险人要掌握保险标的的真实情况有一定的难度。当保险标的的风险增大或基本情况发生变化时，保险人往往难以及时了解，这就成为了保险欺诈存在的可能原因。

2.3.2 核保不严

一些保险公司从业人员素质有待提高，销售人员和核保人员可能为了追求业务量的增加，在承保之前并没有对保险标的的风险进行详细的评估，没有进行严格仔细的核保，对投保方的信用情况也未进行评估，从而导致承保的标的可能风险过高，并未满足一般的承保条件，因此也增加了欺诈发生的可能性。

2.3.3 合同条款局限性

保险条款通常都是保险公司拟定，通过保险合同的签订来规定双方权利义务，约束保险合同双方的行为。在制定保险条款时，保险人会尽量使保险条款能够较好地避免保险欺诈的发生，但由于现实情况往往是复杂多变的，制定保险条款时很难将方方面面的可能性都考虑到，因此很难做到约束所有的欺诈行为。

2.3.4 监管制度不健全

一些保险公司理赔程序不够严谨规范，不容易识别出可疑索赔案件，为保险欺诈提供了可能。还有一些保险公司没有做到有效监管，少数内部从业人员与投保方勾结，骗取保险金，助长了欺诈行为的产生。同时，一些保险公司认为支付保险金的成本低于上法庭的成本，选择支付一些可疑索赔案件，对于欺诈案件，大多保险公司的目的在于拒赔或追回已付赔款，较少追究欺诈行为人的法律责任，这也使得保险欺诈变成了一项低风险高回报的行为，一定程度上助长了保险欺诈。

2.3.5 信息交流有限

因为保险公司之间一般存在的竞争关系，所以很多保险欺诈案件的信息不会在多家保险公司之间进行交流共享，从而导致一家保险公司无法了解在另一家公司有欺诈历史的客户的情况，使得保险欺诈行为的发生更加常见。

2.4 社会原因

2.4.1 法律意识淡薄

进行保险欺诈的投保方在图利时，往往并不了解相关的法律法规，没有意识到欺诈行为的法律后果，并不认为严重的欺诈行为会构成犯罪，而且认为欺诈行为被发现的后果只是退回已得赔款或保险公司拒绝赔付。如果投保方对法律有一定程度上的了解，有强烈的遵纪守法意识，那么就会对自己的行为进行自我约束，法律对人的约束性也就更好地实现了。因此，法律意识淡薄会导致欺诈行为的发生。

2.4.2 社会宽容程度

情节严重的保险欺诈可能构成犯罪，但与其他一些犯罪行为相比，社会公众对于保险欺诈的宽容程度较高。保险欺诈并不属于暴力性的犯罪，虽然会损害保险公司和所有诚实投保方的利益，但对于社会大众来说受害方毕竟使少数，跟自身关系也不密切，而且一些投保方并没有意识到保险欺诈行为会导致保险公司将欺诈增加的成本通过调高保险费来转移，因此会损害到自身的利益，而是认为在保险欺诈案件中，受损的只是保险公司一方，所以社会对于保险欺诈的宽容程度较高，从而也增加了保险欺诈发生的可能。

3 理论基础

随机森林是一种在以决策树为基学习器构建 Bagging 集成的基础上，进一步将随机属性选择引入决策树的训练过程中的集成算法。

决策树是一种常见的机器学习方法，它可以从一组没有规律的事件中通过学习分析和推理，从而对事件进行分类，决策树可以被抽象地理解为一棵树的结构，一般一棵决策树包含了一个根节点，若干个叶节点和若干个内部节点。根节点代表全部的样本，叶节点对应了分类的结果，决策树内部的每一个节点则对应了对一个属性的测试，从根节点到每一个叶节点的路径就对应了一个判定的序列。生成决策树的过程实际上是一个递归的过程，进行决策树学习的目的就是为了得到一棵泛化能力强的决策树。

传统的决策树在进行属性选择时是在节点的所有属性中选出一个最好的属性，而随机森林在进行属性选择时，则是从节点的所有属性中先随机选出一部分属性，然后再从这个子集中选出一个最好的属性。

随机森林的基本单元是决策树，决策树的构造由一个随机向量来决定，随机森林包含了很多个决策树。随机森林先用 bootstrap 重抽样方法在原始样本中抽取出来多个样本，然后针对每个抽取出的样本，用决策树进行建模，得到多个决策树的预测结果后，由多个预测结果共同决定最终的预测结果。

随机森林算法简单并且容易实现，相比 Bagging，随机森林基学习器的多样性不仅仅来自样本扰动，还来自属性扰动，因此进一步提升了最终的泛化性能。

4 模型构建

根据用于建模的数据中因变量是否存在，一般可以将预测分类模型分为两类：有监督学习和无监督学习。有监督学习是进行模型分析的一种主要方法，但有时数据库中不含有作为因变量的数据，这时就需要用到无监督学习方法来进行预测分析，以解决这个限制。

因为在保险业务和索赔相关的数据中，通常没有因变量，即不会有一个明确标记的数据用来表示该索赔是否为可疑索赔，如果数据中存在一个表明是否要求进一步调查的变量，通常可以将该变量作为因变量，如果不存在，则需要用到无监督学习模型。所以，本文选取无监督模式随机森林，对数据进行分类，再用数据库中原本存在的作为因变量的数据对无监督模式的分类结果进行验证，用 R 语言对模型进行实现。

用随机森林进行可疑索赔识别的具体步骤如下：

步骤 1：将索赔数据输入到 R 语言随机森林函数中；

步骤 2：选定树的数量为 500 棵；

步骤 3：运行随机森林函数，得到一个相似度矩阵；

步骤 4：重复步骤 2、3 的过程，以产生第二个相似度矩阵；

步骤 5：取得到的两个相似度矩阵的平均值；

步骤 6：根据计算出的相似度矩阵的平均值，计算差异度矩阵，计算公式为：

$$d_{ij} = \sqrt{1 - p_{ij}} \quad (1)$$

d_{ij}：差异度　　p_{ij}：相似度

步骤 7：对得到的差异度矩阵进行 PAM（Partitioning Around Medoid）聚类；

步骤 8：PAM 聚类将数据分为 k 个簇，分别计算 k=2，3，4 的情况，每一种情况下，每一条记录都会被归入一个特定的组群中；

步骤 9：为了与随机森林进行比较，用欧几里得度量作为差异度进行 PAM 聚类；

步骤 10：将不同集群作为预测器，可疑索赔作为因变量构建一个树模型，用 R 语言 rpart 函数，对不同预测器的重要性进行排序，确定对数据分类效果最好的 k 值，即将数据分成几类。

步骤 11：分别确定随机森林聚类与欧氏距离聚类中，每一类中的可疑索赔比例，对两种方法分类的效果进行比较。

5 实证分析

本文使用的数据是基于 1993 年美国马萨诸塞州汽车保险局数据的模拟数据。原始数据中包含了一百多个变量，1400 个样本。Francis (2016) 基于该数据集和特征工程的结果生成了一个可以公布给大众用于研究的模拟数据集，该数据集共 1500 个样本。

数据集中的变量有两类，一类是索赔记录中存在的典型变量，如医疗机构的数量、类型以及是否有律师参与等；另一类则是代表对索赔是否合理的主观评价的变量。其中有一个用于表示索赔是否可疑的变量 Suspicion，可以用于验证本文模型的分类结果。

数据中一共有 1500 条记录，被标记为可疑索赔的记录有 465 条，约占三分之一。

得到不同集群的变量重要性统计量，从而对不同集群的分类效果进行排序，得到结果如表 1 所示：

表 1 分类效果排名

集群	排名	统计量
随机森林 k=4	1	139.156
随机森林 k=2	2	20.303
随机森林 k=3	3	14.243
欧式距离 k=4	4	9.472
欧式距离 k=3	5	8.323

从排名结果来看，采用随机森林方法，并将所有数据分为 4 类的分类结果最好，能够最好地对是否存在欺诈进行分类、识别。在欧氏距离聚类中，也是 k=4 时，即将所有数据分成 4 类时分类效果最好。下面计算出随机森林聚类和欧氏距离聚类中每一类中的可疑索赔所占的比例，来具体地比较这两种方法的分类效果。

表 2 可疑索赔比例

组别	随机森林		欧式距离	
	可疑索赔占比	数量	可疑索赔占比	数量
1	33%	473	23%	271
2	2%	402	34%	525
3	93%	270	33%	454
4	14%	355	30%	250

从表 2 可以看出，当将所有数据分成四类时，基于随机森林的聚类可疑索赔比例最高的组中，93%为可疑索赔，可疑索赔比例最低的组中仅有 2%为可疑索赔；而基于欧式距离的聚类可疑索赔比例最高的组为 34%可疑索赔，最低的组为 23%可疑索赔，且四组可疑索赔比例都相近，较为平均。所以，随即森林在欺诈识别中的效果更好。

6 研究结论

从对模型的验证可以看出，随即森林聚类分成 4 类时，可以分出一组可疑索赔占比 93%的记录，和一组可疑索赔占比仅 2%，基本都是正常索赔的记录，分类效果很好，用随机森林方法产生差异度再进行聚类，相比于一般的聚类来说，对于可疑索赔的分类效果更好。

本文采用无监督随机森林的方法对数据进行了分类，并且验证了分类结果。研究结果证明，本文使用的一种基于随机森林生成差异度，再进行聚类对索赔数据进行分类从而进行欺诈识别的方法具有可行性。在实际操作中，使用这种方法可以分出欺诈可能性较高的组和欺诈可能性较低的组，分成不同的组别后，就可以对欺诈可能性较高的索赔，在进行理赔时采取更多的措施，给予更多的关注，可以对情况进行更多了解，或者采取进一步的研究与判断，更具有针对性地对保险欺诈进行识别。

参考文献：

[1] 俞昱.保险行业协会反保险欺诈功能研究[D].广东:广东财经大学,2013.

[2] I. L greid. Automatic Fraud Detection-Does it Work Annals of Actuarial Science, 2007,2(2).

[3] Kuo Chung. Lin, Ching Long. Yeh, Shih Ying. Huang. Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance[J]. Applied Mechanics and Materials, 2013,2240(571):1574-1578.

[4] Kose. Ilker, Gokturk. Mehmet, Kilic. Kemal. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance[J]. Applied Soft Computing, 2015,36:283-299.

[5] Johnson. Marina Evrim, Nagarur. Nagen. Multi-stage methodology to detect health insurance claim fraud [J]. Health care management science, 2016,19(3):249-260.

[6] Botond Benedek, Ede László. Identifying Key Fraud Indicators in the Automobile Insurance Industry Using SQL Server Analysis Services[J]. Studia Universitatis Babes-Bolyai Oeconomica, 2019,64(2):53-71.

[7] Hojin Moon, Yuan Pu, Cesarina Ceglia. A Predictive Mode ling for Detecting Fraudulent Automobile Insurance Claims [J]. Theoretical Economics Letters, 2019,9(06):1886-1900.

[8] 杨超.基于 BP 神经网络的健康保险欺诈识别研究[D].山东:青岛大学,2014.

[9] 赵尚梅,赵汀,侯建磊.将支持向量机 SVM 引入机动车保险欺诈识别[J].中国保险, 2015,(8):15-19.

[10] 李聪.中国健康保险欺诈的理论分析与实证研究[D].山东:青岛大学,2015.

[11] 裴晨.基于随机森林与 GBDT 的社会医疗保险欺诈识别问题研究.辽宁:东北财经大学,2018.

[12] 曹鲁慧,秦丰林,闫中敏.基于 TLSTM 的医疗保险欺诈检测[J/OL].计算机工程与应用:1-7[2020-04-18].