

数据挖掘技术在保险客户信用评估的应用

陈艳, 张燕平

(安徽大学人工智能研究所 智能计算与信号处理重点实验室, 安徽 合肥 230039)

摘要: 目前, 数据挖掘技术广泛应用于各个领域。文中将数据挖掘应用于保险客户在信用等级分类中, 即采用了基于神经网络的覆盖算法作为客户信用评分分类器的设计算法。通过对保险数据的分析, 对保险用户信用等级进行分类, 降低了人为因素的评价干扰。通过分类实验表明, 覆盖算法的准确性和网络训练速度都大大高于 SVM。为保险公司有针对性的调查提供了一定的参考依据。

关键词: 数据挖掘; 覆盖算法; 保险; 信用评估

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2008)05-0179-03

Application of Data Mining in Credit Sorting of Insurance Client

CHEN Yan, ZHANG Yan-ping

(Key Lab. of Intelligent Computing & Signal Processing, Inst. of Artificial Intelligence,
Anhui Univ., Hefei 230039, China)

Abstract: Recently, data mining technology is widely applied in many fields. In this paper, data mining is applied to the credit sorting of insurance client, which used the covering algorithm on which based neural networks as the classification of clients' credit. By analyzing the insurance data set, the algorithm sort the clients' credit, man-made disturbances are greatly eliminated. The experiment showed that the accuracy and speed of the covering algorithm is better than that of SVM. Then, it provides certain consultation for insurance company on farther investigating.

Key words: data mining; covering algorithm; insurance; credit sorting

0 引言

近年来, 随着人们消费水平的不断提高和消费意识的改变, 越来越多的人开始购买各式各样的保险, 以对日后不测之事做好准备, 加强保障。与此同时, 保险行业随之迅速发展。在保险行业获取利益的同时也注意到保险行业存在的巨大风险, 因此对保险欺诈防范和识别成为近来研究的热点之一。目前, 数据挖掘技术已经广泛应用于商业决策中^[1-4]。在保险行业中, 将数据挖掘技术用于已有巨大信息量的客户保险数据库, 对客户风险进行预测, 将客户信用分成若干个等级, 从而有针对性地对信用等级预测低的个别客户

进行详细的调查核实材料, 可以有效地避免客户骗保的发生, 降低公司保险投资风险, 尽量避免经济损失。

信用评分实际上是将一个总体按照不同的特征分成若干个不同组的一种方法, 即是一种数据分类方法。目前分类方法主要包括基于决策树分类方法、基于神经网络分类方法和基于统计的分类方法。文中使用基于神经网络的覆盖算法对保险数据进行分析分类。通过对某个保险公司某个地区的寿险中的大量数据进行处理、提取、挖掘, 将参保客户按信用度分类, 为保险公司有针对性的调查提供了一定的参考依据。

1 覆盖算法简介

张铃、张钹教授在文献[5]中给出了 M-P 神经元的几何意义, 并在文献[6]中提出了神经网络交叉覆盖设计算法。

1.1 M-P 神经元的几何意义

算法中首先假定样本输入向量都落在 $n+1$ 维空间的某个超球面 S^n 上(球面半径不妨设为 R), 否则, 可通过变换: $T(x) = [x, \sqrt{(r^2 - |x|^2)}]$ 将样本点映

收稿日期: 2007-08-19

基金项目: 国家自然科学基金(60475107; 60675031); 973 计划资助项目(2004CB318108); 安徽省教育厅重点自然科学基金(2006KJ015A; 2005kj053); 安徽省自然科学基金(0504200208); 安徽大学 211 工程学术创新团队资助项目

作者简介: 陈艳(1983-), 女, 硕士研究生, 研究方向为数据挖掘与智能计算; 张燕平, 教授, 硕士生导师, 研究方向为人工智能、人工神经网络的理论与应用。

射到球面 S^n 上,其中 $r \geq \max\{\|x\| \mid x \in D\}$ 。那么这时 $W * (x - \theta) > 0$ 就表示由超平面 P 所分割的正半空间部分。其中: W 是权向量, θ 是阈值, x 是样本输入向量。正半空间的部分恰为超球面上的某个“球形领域”,若 W 与 x 等长,则 W 为这个球形领域的中心,其半径为 $r(\theta) = R \arccos(\theta/R)^2$,若取神经元的功能函数为符号函数 $\text{Sgn}(W * x - \theta)$,取其激励函数为 $\text{Sgn}(W * x - \theta) = \begin{cases} 1, & x > 0 \\ 0, & \text{其它} \end{cases}$,则其功能函数为超球面上“球形领域”的特征函数。这样,一个神经元就对应超球面上一个“球形领域”^[5]。

由神经元的几何意义得知,构造一个网络,对给定的样本集能够进行符合要求的分类,等价于求出一组覆盖领域,这组覆盖领域可以将样本按要求分隔开领域。这样,就将神经网络的最优设计问题转化成某种求最优覆盖的问题。

1.2 覆盖算法

覆盖算法^[6]的基本思想就是用求出的覆盖领域作为三层网络的隐含层,输入层为测试集,输出层为测试集的分类结果。给定一输入集 $K = \{x^1, x^2, \dots, x^k\}$,设 K 分为 s 个子集 $K^1 = \{x^1, x^2, \dots, x^{m(1)}\}, \dots, K^s = \{x^{m(s-1)+1}, x^{m(s-1)+2}, \dots, x^k\}$ 。用一个三层网络构造分类器,求出一组领域,这组领域将不同类的点分隔开来,使属于 K^i 的点的输出均为 $y_i = \{0, \dots, 0, 1, 0, \dots, 0\}$ (即其第 i 个分量为 1,其余分量为 0 的向量), $i = 1, 2, \dots, s$ 。

算法步骤如下:

第一步,找初始点 a ,从点 a 开始覆盖。

第二步,确定以 a 为圆心的覆盖领域 C^1 的半径。找离 a 最近的异类点,其距离记 d_1 ;找离 a 最远的距离小于 d 的同类点,其距离记为 d_2 。覆盖领域的半径为 $r = (d_1 + d_2)/2$ 。

第三步,求领域 C^1 的重心,并记为 a 。

第四步,重复第二到第三步,直到覆盖的点不再增加为止。这样就得到一个覆盖 K^1 中点的局部最大覆盖 C^1 。其覆盖 K^1 记为 K^{1i} 。

第五步,找一个不同类点 a 再开始覆盖;令 $T \leftarrow K^1/K^{1i}, K^1 \leftarrow K^2, K^2 \leftarrow T$ 。

若只剩最后一类点,则将最后一类点作为一个覆盖;否则回第二步。

2 基于交叉覆盖算法的客户信用分类

2.1 数据准备

文中所用的数据来自于某保险公司的寿险数据库,其中包括数十张数据表。通过分析,确定与本次挖

掘相关的数据表。初步得到的数据表,信息量大,并且维数多,表中有些数据并不完备,有些字段与挖掘无关。为了解决这些问题,采用了多表合并的抽取方法,使用 SQL 查询语句,根据表中的主键、外键将这些数据表逐步合并成一张数据表,并获取适合需求的字段。

2.2 数据预处理

初步采集后的数据并不完整,需要对其进行一定的预处理^[7],才能满足挖掘算法输入的要求。初步采集得到的合并的数据表,经过空缺值的处理、冗余属性的去除、数据归一化和离散化操作,得到的数据样本情况如表 1 所示。

表 1 样本概况

样本数量	13690
样本维数	14

根据原始数据中的客户黑名单登记表,查找出总表中曾经发生过骗保或拒赔的客户。其中黑名单样本数量是 1341 条,将近是未发生骗保事件样本数目的十分之一。将客户信用等级作为决策属性,可将这些样本分为 2 类:一类未出现在黑名单中,信用良,另一类出现在黑名单中,信用差;也可根据客户黑名单表中数据的不同情况将样本分为 3 类,即将出现在黑名单中的数据根据骗保和拒赔分为 2 类:信用差,信用极差。样本分布情况如表 2 所示。

表 2 样本分布情况

样本类别		样本数目
分两类	第一类	12349 条
	第二类	1341 条
分三类	第一类	12349 条
	第二类	439 条
	第三类	908 条

2.3 实验结果与分析

基于交叉覆盖算法的信用分类算法的实质就是通过对样本的学习构造神经网络,产生分类器,再通过分类器对样本进行分类。将样本分为 10 组,9 组作为训练,1 组作为测试,循环交叉。测试神经网络的分类效果,分类准确率由如下公式求出:

$$\text{分类准确率} = \frac{\text{正确分类样本数}}{\text{测试样本总数}} \times 100\%$$

由于上述不同类别的数据数量差别较多,用随机方法分别从不同类别的数据中抽取部分数据作为实验数据,数据情况如表 3、表 4 所示。

实验采用了交叉覆盖算法和 SVM 方法^[8]对样本分 3 类的情况进行了测试,结果如表 5、表 6 所示。

从表 5、表 6 可以看出交叉覆盖算法和 SVM 法的分类准确率随着样本数目的增多有所提高,但是交叉

表3 1类样本数目是2类样本数目5倍

样本类别		样本数目
分两类	第一类	5500 条
	第二类	1100 条
分三类	第一类	5000 条
	第二类	374 条
	第三类	726 条

表4 1类样本是2类样本数目3倍

样本类别		样本数目
分两类	第一类	3200 条
	第二类	1100 条
分三类	第一类	3000 条
	第二类	374 条
	第三类	726 条

表5 对表3实验结果

学习样本个数	交叉覆盖算法		SVM法	
	分类准确率	训练时间	分类准确率	训练时间
1000	75.8%	0.4305	59.4%	0.8740
2000	80.3%	0.7580	62.3%	1.9730
4000	86.6%	1.1620	65.5%	3.5610
6000	87.7%	1.6340	67.6%	6.0810

表6 对表4实验结果

学习样本个数	交叉覆盖算法		SVM法	
	分类准确率	训练时间	分类准确率	训练时间
1000	71.7%	0.4190	57.9%	0.7920
1500	79.6%	0.6960	61.1%	1.7310
3000	85.4%	0.9690	63.5%	3.0280

覆盖算法的准确率和训练时间都更好。尤其当样本数量越多时,SVM分类器训练时间的增幅要明显高过覆盖算法。在从表5与表6的对比中看出,当信用度低的记录减少时,算法的准确性会小幅度提高。这对于保险业中不良客户的比例也很小这个特点来说,覆盖网络的分类准确性能够得到保证。

(上接第178页)

的,它几乎涉及到了当今计算机科学的各个领域。无论是网络技术还是多媒体技术的发展都会促使它变得更加完善。就它的应用前景来说,它不仅可应用于学校,还能应用于会议中心、培训机构、企业等单位。

参考文献:

- [1] Steinmetx R, Nahrstedt K. 多媒体技术:计算、通信和应用[M]. 潘志庚,叶 绿,耿卫东等译. 北京:清华大学出版社,2000.
- [2] 罗万伯. 多媒体技术网络课程(新世纪网络课程建设工程). [M]. 北京:高等教育出版社,高等教育电子音像出版社,2004.
- [3] Andleigh P K, Thakrar K. 多媒体系统设计[M]. 徐光佑,史

3 结束语

将数据挖掘技术应用到保险客户信用度评分上,在一定程度上为商业决策提供了一个客观的参考依据。

从实验结果看来,分类的准确性最高达到了87.7%,精度高于SVM分类方法,并且训练时间短,网络建立速度快。由于数据集中的数据并不完备,若对不完备数据做进一步的有效处理,并不是如文中直接删除,应该能够进一步提高网络分类准确率。另外,寿险方面有很多其他影响客户信用的因素,并未出现在数据集中,这一点也影响了分类的准确性。如何提取数据表中的数据,进行更有效的处理,提高分类的准确性,还需要更深入的研究和实验。

参考文献:

- [1] 邓乃扬,田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京:科学出版社,2004.
- [2] 侯惠芳,刘素华. 基于支持向量机的商业银行信用风险评估[J]. 计算机工程与应用,2004,40(31):176-192.
- [3] 胡光杰. 数据挖掘在供应商评价中的应用[D]. 合肥:安徽大学,2006.
- [4] 谢友辉,蒋新华. 数据挖掘技术及在保险领域中的应用[J]. 信息技术,2003,27(8):5-11.
- [5] 张 铃,张 钺. M-P 神经元模型的几何意义及其应用[J]. 软件学报,1998,9(5):334-338.
- [6] Zhang Ling, Zhang Bo. A Geometrical Representation of McCulloch-Pitts Neural Model and Its Applications[J]. IEEE Trans. on Neural Networks, 1999,10(4):925-929.
- [7] 霍凌慧,马少平,唐焕玲. 银行卡分类挖掘数据的预处理[J]. 计算机工程,2003,29(11):195-197.
- [8] 张 铃. 基于核函数的SVM机与三层前向神经网络的关系[J]. 计算机学报,2002,25(7):696-699.

元春译. 北京:电子工业出版社,1998.

- [4] Comer D E. Internetworking with TCP/IP Volume I[M]. 3rd ed. [s.l.]:Prentice Hall, 1998.
- [5] Tanenbaum A S. Computer Networks[M]. 3rd ed. [s.l.]: Prentice Hall, 1997.
- [6] 周 丹,商卫东,肖 菁. 动态媒体流的同步实现技术[J]. 武汉大学学报,2000,46(1):46-48.
- [7] Battista S, Casalino F, Lande C. MPEG-4: a multimedia standard for the third millennium[J]. Multimedia, IEEE, 1999,6(4):74-83.
- [8] Schuzrinne H, Casner S, Frederick R, et al. RTP: A Transport Protocol for Real-Time Application[S]. RFC1889. [s.l.]: Internet Engineering Task Force, 1996.