

稀有类分类问题探讨

职为梅, 范 明

(郑州大学 信息工程学院, 河南 郑州 450052)

摘 要:分类是数据挖掘中的重要任务之一,稀有类分类问题是分类中的一个重要分支,可以描述为从一个分布极不平衡的数据集中标识出那些具有显著意义却很少发生的实例,在现实生活中的很多领域都有广泛的应用。详细地介绍了稀有类分类的问题,探讨了稀有类分类的一些特征、影响稀有类分类的一些因素和对稀有类分类进行评估的标准,介绍了当前分类稀有类的主要方法:基于数据集的方法和基于算法的方法。介绍了当前几种流行的稀有类分类算法。

关键词:分类;稀有类;显露模式;两阶段分类

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2010)07-0250-04

Research on Classification of Rare Classes

ZHI Wei-mei, FAN Ming

(College of Information Engineering, Zhengzhou University, Zhengzhou 450052, China)

Abstract: Classification is an important task in data mining. Rare classification is a part of classification and it can be described as identifying the instance with statistical significance from imbalanced datasets. The classification of rarely occurring cases is widely used in many real life applications. Introduce the question of rare classification and discuss the features and general criteria of rare classification, and also study the popular methods to classify rare cases: based on data level and algorithm level. In the last introduce the popular algorithm of rare classification.

Key words: classification; rare class; emerging pattern; two-phase classification

0 引 言

分类是数据挖掘中的重要任务之一,在商业、金融、电讯、DNA分析、科学研究等诸多领域具有广泛的应用。统计学、机器学习、神经网络等领域的研究者提出了很多分类方法^[1]。分类稀有类是分类中的一个重要问题。这个问题可以描述为从一个分布极不平衡的数据集中标识出那些具有显著意义却很少发生的实例。分类稀有类在现实生活中的很多领域都有广泛的应用,例如,网络侵入检测、欺骗探测和偏差探测。在网络入侵中,一个计算机通过猜测一个密码或打开一个ftp数据连接进行远程攻击。虽然这种网络行为是不常见的,但识别并分析出这种行为对于网络安全却是很有必要的。

由于稀有类实例数目很少,很难提供很完备的信息,使得分类稀有类问题变得更具有挑战性。传统的分类算法在分类稀有类时往往失效,因为分类器是建

立在训练数据集上,得出最适合数据集特征的结果,训练数据集中稀有类的实例数量少,使得分类器不倾向于稀有类,分类准确率低。文中将详细分析稀有类分类的特征、稀有类的评估标准以及当前分类稀有类的一些方法。

1 稀有类分类介绍

普通分类问题中,各个类包含的数据分布比较平衡,稀有类分类问题中,数据的分布极不平衡。例如:将一批医疗数据分类为“癌症患者”和“非癌症患者”两个类,其中“癌症患者”是小比例样本(假设占总样本的1%),称其为目标类,“非癌症患者”为多数类样本,称为非目标类,从大量数据中正确识别“癌症患者”就是稀有类分类问题。由于在数据集中所占比率太小,使得稀有类分类问题比普通分类问题更具挑战性。

稀有类问题有三个显著特征:

(1)稀有性:这是稀有类分类问题的重要特征,正是因为目标类的稀有性使得稀有类分类问题难于普通的分类问题,如何快速、准确地识别稀有类并避免分类器对训练数据集的过分拟合是目前极具挑战性的问

收稿日期:2009-10-19;修回日期:2010-01-28

基金项目:河南省自然科学基金(0211050100)

作者简介:职为梅(1977-),女,讲师,硕士研究生,从事数据挖掘的研究。

题。

(2)应用性:稀有类分类在现实生活的很多领域都有广泛的应用,比如:网络入侵、欺诈检测、疾病诊断等。例如网络入侵,大多时候都是正常的连接访问,偶尔情况下是会有黑客攻击,尽管不经常发生,但识别出极少情况下的攻击行为对网络的安全至关重要。

(3)多态性:通常,每个目标类 C 和非目标类 NC 又包括多个子类,每个子类又有不同的特征。从非目标类 NC 中区分目标类 C 转化为从 NC 的多个子类中区分 C 类及其子类,这就形成多个子类之间的多态性,导致问题复杂化^[2]。

2 影响稀有类分类的因素

通常认为影响稀有类分类的因素是不平衡的类分布(Imbalanced class distribution),但是通过大量的研究和实验证明,数据的不平衡性只是影响稀有类分类的一个因素,还有两个重要的因素影响稀有类分布:小样本规格(Small sample size)、分离性(Separability)和类内子概念的存在(Within-class sub-concepts)。下面简单讨论这四个因素对稀有类分类的影响。

不平衡的类分布:研究表明,类分布越是相对平衡的数据分类的性能越好^[3]。探讨了训练集的类分布和判定树分类性能的关系,但是不能确定多大的类分布比率使得分类性能下降。研究表明,在有些应用中 1:35 的时候不能很好地建立分类器,而有的应用中 1:10 就很难建立了。

小样本规格:给定特定的类分布比率,样本大小在确定一个好的分类模型中起着非常重要的作用,对于有限的样本发现稀有类内在的规律是不可能的。研究表明,随着训练样本增加由不平衡类分布造成的错误降低,这主要因为稀有类实例越多,可获得的关于稀有类的信息就越多。

分离性:从普通类中区分出稀有类是稀有类分类的关键问题。假定每个类中存在高度可区分模式,则不需要很复杂的规则区分它们。但是如果有一些特征空间上不同类的模式有重叠就会极大降低被正确识别的稀有类实例数目。

类内子概念的存在:在许多分类问题中,单一类通常由许多子概念构成,类实例由多个子类的类实例构成,各个子类并不总是包含相同的类实例。与类间概念的分布不平衡相对应,这种现象被称为类内不平衡(within-class imbalance)。类内不平衡性使得分类效果更差,主要表现在以下两方面:(1)类内子概念的存在使得数据集更加复杂;(2)在多数情况下类内概念不清楚。

根据以上分析,由于影响稀有类分类的因素多种多样,使得稀有类分类问题更加复杂,分类的性能降低。

3 稀有类分类的评估标准

常用的分类算法的评估标准有:预测的准确率、速度、强壮性、可规模性及可解释性。对于普通类来说,通常使用分类器的总准确率来评价分类效果。然而对于稀有类分类问题来说,由于关注的焦点不同,仅用准确率是不合适的。

例如:将一批医疗数据分类为“癌症患者”和“非癌症患者”,总体上分类准确率为 99%,这看似相当准确,但如果实际上只有 1% 的训练样本是“癌症患者”呢?预测所有实例都是“非癌症患者”依然可以达到该准确率,然而对于目标类“癌症患者”来说,效果却极差。故而在稀有类分类问题中更关注稀少目标类的正确分类率。在评价稀有类分类时,还应该采用其他的评价标准。

这里假设只考虑包含两个类的二元分类问题,假设 C 类为目标类,即稀有类,NC 为非目标类,根据分类器的预测类标号和实际类标号的分布情况存在如表 1 所示的混合矩阵(Confusion Matrix)。

表 1 二元分类问题的混合矩阵

| | 预测为 C 类 | 预测为 NC 类 |
|----------|---------|----------|
| 实际为 C 类 | TP | FN |
| 实际为 NC 类 | FP | TN |

根据表 1 得到如下度量:

$$TPrate = \frac{TP}{TP + FN};$$

$$TNrate = \frac{TN}{TN + FP};$$

$$FPrate = \frac{FP}{TN + FP};$$

$$FNrate = \frac{FN}{TP + FN};$$

$$PPvalue = \frac{TP}{TP + FP};$$

$$NPvalue = \frac{TN}{TN + FN}$$

最理想情况下:TPrate = 1, FPrate = 0, 分类准确率为 100%;如果 TPrate = 0, FPrate = 0, 所有类实例被划分为非目标类,如果 TPrate = 1, FPrate = 1, 所有类实例被划分为目标类。通常情况下使用召回率(recall)即 TPrate、精确率(precision)即 PPvalue 和 F-度量来评估稀有类分类。

$$F-度量(F-measure)由下式定义:F = \frac{2RP}{R + P}$$

其中 R 为 recall, P 为 precision。

4 稀有类分类方法讨论

判定树、支持向量集、神经网络、贝叶斯网络、最邻近算法等传统的分类模型以及新提出的分类模型在分类稀有类时失效,使得大量的专家学者对稀有类分类问题进行了详尽的研究,研究表明,解决稀有类分类问题的方法总体上可以分为两类:基于数据集的和基于算法的^[4]。基于数据的思想如下:通过重新取样使得分布不平衡的数据集变成平衡数据集,包括增加稀有类实例或者减少普通类实例。基于算法的思想如下:通过改进已有分类算法使得它更倾向于稀有类,例如 cost-sensitive learning。

目前稀有类分类算法比较少,主要有 Ramesh Agarwal 和 Mahesh V. Joshi 提出的 PNrule 方法(即两阶段规则归纳算法)^[5]、Hamad Alhammady 等提出的 EPRC 算法^[6]、我们的一些工作提出的 CREEPTP 分类方法^[7](即利用基本显露模式两阶段分类稀有类);一些传统的分类算法(如 NB 和 C5.0)虽然也可以用来分类稀有类,但效果并不理想,还有使用集成分类器分类 Bagging、Boost: AdaBoost 等。

4.1 基于数据集的方法

基于数据集的方法有很多,比如,随机增加稀有类实例数据(Over-Sampling)、随机减少普通类实例数据(Under-Sampling)、有指导地增加稀有类实例数据、有指导地减少普通类实例数据、通过合成新的数据增加稀有类实例数据,或者是上述方法的综合。

实际处理稀有类问题时,使用最多的就是随机取样增加稀有类实例数据或者随机减少普通类实例数据。但这种方法存在一个问题,对于一个特定的数据集来说什么样的类分布是最好的?通常认为对于两个类,类分布比率 1:1 最好,但不同的数据集并不一样。对于随机取样还存在一些问题,比如,如何有效的取样?

4.2 基于算法的方法

基于算法的方法是通过调整现有算法或者提出新的分类方法在分类时有利于稀有类。比如 R. Agarwal 和 M. V. Joshi 提出的“基于规则的两阶段分类稀有类”方法(即 PNrule)、Hamad Alhammady 等提出的 EPRC 算法以及我们的一些工作提出的 CREEPTP 分类方法(即利用基本显露模式两阶段分类稀有类)。

两阶段分类方法将训练过程分为两个阶段,第一个阶段从整个数据集开始训练规则,采用顺序覆盖技术迭带产生规则,使训练得到的规则尽量覆盖多的正例(目标类实例),不过分考虑其覆盖的反例(非目标类实例),该阶段训练得到的规则称为 P 规则,第一个阶段追求高覆盖率,尽可能多地覆盖稀有类数据;第二个

阶段从所有的 P 规则覆盖的实例集开始,P 规则覆盖的正例和反例分别变为第二阶段的反例和正例,训练得到的规则称为 N 规则,第二个阶段追求高精确率,尽可能多地删除普通类实例。第二阶段的存在使第一阶段对有错误倾向的小覆盖问题^[8]不太敏感,在第一个阶段中,可以尽可能地覆盖更多的正例,而不用对其准确度做太多的考虑,因为在第二阶段中训练规则会去除反例。这就是前面提到的在第一阶段中取得高覆盖。第二阶段有很好的能力获得目标类不存在的特性,因为它将所有的反例连接起来,这使两阶段方法对反例分离问题不太敏感。注意由第二阶段造成的不同在于 PN 规则需要在该阶段获得非目标类存在的特性,这与 Ripper 和 C4.5 不同,它们需要获得目标类不存在性的特性以区别非目标类。由于在第二阶段学习 N 规则去除反例,故可以提高规则的精确度。

由上面的分析可以得出前面的结论:两阶段方法能够解决绝大多数顺序覆盖技术所不能解决的问题,小覆盖问题和反例分离问题,并且能够在两个阶段中分别获得高覆盖和高精确度。两个阶段相结合从而在稀有类分类上具有很强的优势。基于规则的两阶段分类方法很好地克服了传统的分类方法(如 C4.5 和 Ripper)在稀有类分类时遇到的两个问题:反例碎片问题和有错误倾向的小覆盖问题。实验结果表明,基于规则的两阶段分类法具有较好的分类效果,特别适合对稀有类进行分类,其分类的误差和对稀有类的误分类率都显著低于 C4.5 和 Ripper。

EPRC 算法使用显露模式^[9,10] EP(Emerging Patterns)来分类稀有类。EP 是一个项集(项的集合),其支持度由一个类到另一个类显著增加。两个支持度的比称作 EP 的增长率。例如,假定有顾客数据集,包含类 buys-computer = “yes”或 C1 和 buys-computer = “no”或 C2。项集 {age = “<=30”, students = “no”} 是一个典型的 EP,其支持度由在 C1 中的 0.2% 增长到在 C2 中的 57.6%,增长率为 288(即 57.6/0.2)。注意,一个项或者是分类属性上的简单相等测试,或者是检查数值属性是否在某个区间的测试。每个 EP 是一个多属性上的测试,并且可能在区分一个类的实例与另一个类的实例方面非常强。例如,如果一个新样本 X 包含在上面的 EP 中,可以说 X 属于 C2 的几率为 99.6%。一般地,EP 的区分能力大约正比于它的增长率和它在目标类的支持度。显露模式是那些支持度从一个数据集到另一个数据集发生显著变化的项集,这些项集能够捕获数据库中的显露趋势,因此能够很好地用来作为分类基础。EPRC 算法从训练数据集中挖掘 EPs,根据训练出来的 EPs 构建稀有类实例,从而增

加有意义的稀有类实例数目,构建分类器分类。实验表明 EPRC 算法能够有效识别稀有类,提高分类的性能。

2005 年笔者提出了利用基本显露模式 eEP (essential Emerging Patterns) 的两阶段分类稀有类方法 (即 CREEPTP)。基本显露模式 eEP 是一种特殊的 EP,在稠密的数据集中通常包含大量的 EP,实践表明使用大量 EP 分类并不是很有效。H. Fan 和 K. Ramamohanarao 提出使用一种特殊的 EP, eEP (essential Emerging Patterns) 进行分类^[11]。eEP 是 EP 边界表示的左边界的子集^[12],具有很高的增长率。eEP 去掉了那些包含冗余信息和噪声的项集,其数量比 EP 少得多,并且具有很高的分类效率。CREEPTP 使用 eEP 作为分类的基础,并采用两个阶段来挖掘基本显露模式 eEP,第二个阶段作为第一个阶段的纠正来提高稀有类分类的精度。使用评价稀有类分类的两个常用尺度:召回率 (Recall) 和精度 (Precision)。实验结果表明该算法在分类稀有类的时候可以取得较好的召回率和精度。实验结果和几个经典的分类算法比较,在分类准确性上也达到了很好的性能。

5 结束语

文中针对稀有类分类问题进行了研究。文章给出了稀有类分类问题的概念,并分析了稀有类分类问题的特征,从而给出稀有类分类问题的难点所在。针对稀有类问题,给出探讨了评估标准以及当前分类稀有类的主要方法。

参考文献:

- [1] Han J, Kanber M. 数据挖掘:概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2001.
- [2] 刘艳霞,职为梅,杨亮. 稀有类分类问题研究[J]. 微型机与应用,2005(6):54-56.
- [3] Weiss G, Provost F. Learning when training data are costly:

 (上接第 131 页)

1980.
- [2] 韩东海,王超,李群. 入侵检测系统实例剖析[M]. 北京:清华大学出版社,2004.
- [3] 邵峰晶,于忠清. 数据挖掘原理与算法[M]. 北京:中国水利水电出版社,2003.
- [4] 熊忠阳,周亚峰. Web 访问挖掘的预处理技术的研究[J]. 计算机技术与发展,2007,17(8):11-14.
- [5] 孙吉贵,刘洁,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.
- [6] 韩家炜,堪博. 数据挖掘:概念与技术.[M]. 第2版. 范明,孟小峰,译. 北京:机械工业出版社,2007.
- [7] Li Lingjuan, Tang Wenyu, Wang Ruchuan. A CBR Engine

the effect of class distribution on tree induction[J]. J. Artif. Intell. Res., 2003, 19: 315-354.

- [4] Yan Min, Kamel M S, Wong A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007(10): 3358-3378.
- [5] Agarwal R, Joshi M V. PNrule: A new Framework for Learning Classifier Models in Data Mining (A Case-Study in Network Intrusion Detection) [C]//Proc. of the First SIAM Conference on Data Mining. Chicago, USA: [s. n.], 2001.
- [6] Alhammady H, Ramamohanarao K. The Application of Emerging Patterns for Improving the Quality of Rare-class Classification[C]//Proc. of the 8th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD2004). Sydney, Australia: [s. n.], 2004: 207-211.
- [7] 职为梅,范明. 利用基本显露模式两阶段分类稀有类[J]. 微机发展(现更名:计算机技术与发展), 2005, 15(12): 44-47.
- [8] Agarwal R, Joshi M V, Kumar V. Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction[C]//the Proc of ACM SIGMOD/PODS. [s. l.]: [s. n.], 2001: 91-102.
- [9] Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules[C]//ICDM'01. San Jose, CA: [s. n.], 2001: 369-376.
- [10] Dong G, Zhang X, Wong L, et al. CAEP: Classification by Aggregating emerging patterns[C]//Proc. of the 2nd Int'l Conf. on Discovery Science (DS'99). Tokyo, Japan: [s. n.], 1999: 30-42.
- [11] Fan H, Ramamohanarao K. A Bayesian Approach to use Emerging Patterns for Classification[C]//Proc of 14th Australasian Database Conference. Australia: Australian Computer Society, Inc, 2003: 39-48.
- [12] Dong G, Li J. Efficient mining of emerging patterns: Discovering trends and differences[C]//Proc. of KDD'99. San Diego, USA: [s. n.], 1999: 15-18.
- Adapting to IDS[J]. Lecture Notes on Artificial Intelligence, 2005, 3802: 334-339.
- [8] MIT. MIT's KDD Cup 99 dataset[EB/OL]. 1999-10. <http://kdd.ics.uci.edu/databases/kddcup99.html>.
- [9] 李玲娟,梁玉龙,王汝传. 适用于IDS中数据分类的数值归约算法[J]. 计算机应用研究, 2007, 24(12): 146-148.
- [10] 周桂芳. IDS中特征选择算法的研究[D]. 南京:南京邮电大学, 2006.
- [11] 刘汉良. 统计学教程[M]. 第3版. 上海:上海财经大学出版社, 2005.
- [12] 梁玉龙. 数据挖掘技术及其应用研究[D]. 南京:南京邮电大学, 2008.