基于聚类和Ripper的稀有类分类方法

余 雯[1], 蒋盛益[1], 黄兴全[2]
(1. 广东外语外贸大学信息学院, 广东 广州 510006; 2. 广东蓝鸽科技有限公司, 广东 广州 510540)

[摘 要] 稀有类分类在许多领域有重要应用,针对稀有类在数据中所占比例少,容易被忽略的特点,提出一种基于聚类和Ripper的稀有类分类方法,该方法在一趟聚类的结果中,通过将在整个数据集中所占的比例低于15%的聚类标识为少数类,再应用Ripper分类算法分别对少数类和多数类分别进行分类建模,并按照一定的组合方式调整得出整个数据集的最终规则集. 在UCI数据集上的测试结果表明,基于一趟聚类和Ripper的稀有类分类方法对稀有类可产生高质量的分类效果. 可以将该方法应用于现实生活的领域中进行稀有数据的分类.

[关键词] 数据挖掘; 稀有类分类; 一趟聚类
[中图分类号] TP311.12 [文献标识码] A [文章编号] 1000-9965(2009)01-0040-05

A rare-class classification approach based on Clustering and Ripper

YU Wen[1], JIANG Sheng-yi[1], HUANG Xing-quan[2]
(1. College of Information, Guangdong University of Foreign Studies, Guangzhou 510006, China;
2. Guangdong Lancoo Co, Limited, Guangzhou 510540, China)

[Abstract] The rare-class classification is an important issue in many real life applications; this paper considers the rare-class datasets are easily ignored in the classification because of its low proportion of the whole datasets. We apply a rare-class classification approach based on clustering and Ripper. This approach is trying to find out the rare-class datasets after Cluster through recognizing every cluster whose proportion of the whole datasets is lower than 15% as the rare-class datasets. After that, Ripper algorithm is used to classify both the rare-class datasets and the normal-class datasets separately. The rule set of the whole datasets will be created by the certain method of this approach according to the model which has already been set up above. The experiments carried on benchmark datasets from the UCI Machine Learning Repository show that this approach creates high quality classifying. This approach can also be implemented to classify the rare-class datasets in some practical life applications.

[Key words] data mining; rare-class; classification; One-pass Clustering

在模式识别的实际分类应用中,经常碰到类分布不平衡问题,而且是多分类的不平衡分布问题. 如医疗检查、产品质量检测、网络入侵检测等,其中合格或正常的是多数类,少数类或弱势类虽然数量少,

[收稿日期] 2008-09-25
[基金项目] 国家自然科学基金项目(60673191);广东省高等学校自然科学研究重点项目(06Z012);
广东外语外贸大学科研创新团队项目(GW2006-TA-005)
[作者简介] 余 雯(1987-),女,硕士研究生. 研究方向:数据挖掘;通讯作者:蒋盛益,教授

但对这些少数类的识别却非常重要,在许多情况下需要对这些稀有类建立分类器,以确定未知的稀有类对象属于哪个预定义的目标类.这种类别不平衡问题,一方面表现为不同类的样本数量和先验概率的很大不同,一般为1:10以上,甚至1:10 000;另一方面表现为错分代价的不平衡,往往是少数类被错分为多数类的代价要大得多.在不平衡类分布下,传统的分类算法不能很好地解决这一问题,因为传统的分类器通常把具有重要潜在价值的稀有类当成噪声剔除来进行分类,而且是以总体分类精度为评价指标.例如网络入侵一般概率为1%以下,绝大多数为正常访问,如果检测识别全部为正常访问,精度也有99%,但显然这一精度对于网络入侵这一稀有类分类的评估是毫无意义的.由于稀有类问题[1-4]的特殊性、复杂性及难解性,很难使用传统的分类器对它们进行准确分类.

目前在稀有类分类方面已有许多研究工作,国外比较典型的有Ramesh Agarwal和Mahesh V. Joshi提出的PNrule方法[3-4]及Hamad Alhammady和Kotagiri Ramamohanarao[5]提出的EPRC算法以及集成分类方法.国内比较典型的有刘艳霞、范明等[6]提出的基于EEP的稀有类分类方法,和范明、职为梅[7]提出的利用基本显露模式的两阶段稀有类分类方法.现有稀有类分类方法在准确性、时效性以及把有重要潜在价值的异常点当成噪声被剔除等方面尚不能得到完善的解决.针对目前稀有类分类算法在稀有类分类问题上存在的不足,本文提出基于一趟聚类和Ripper的稀有类分类方法.

1　一趟聚类算法和Ripper分类算法简介

本文采用文献[8]中簇的表示、距离定义及一趟聚类算法,一趟聚类采用最小距离原则对数据进行聚类,具体过程如下:

(1)初始时,聚类集合为空;读入一个新的对象;

(2)以这个对象构造一个新的簇;

(3)若已到数据库末尾,则转(6),否则读入新对象,利用给定的距离定义,计算它与每个已有簇间的距离,并选择最小的距离;

(4)若最小距离超过给定的阈值d,转(2);

(5)否则将该对象并入具有最小距离的簇中并更新该簇的各分类属性值的统计频度及数值属性的质心,转(3);

(6)结束.

Ripper分类算法[9]是一种基于规则的分类器,对两分类问题,Ripper算法选择多数类作为默认类,并为预测少数类学习规则;对于多类问题,先按类的频率对类进行排序,设$\{C_1, C_2, \cdots, C_k\}$是排序后的类,其中$C_1$是最不频繁的类,而$C_k$是最频繁的类,在第一次迭代中,把属于$C_1$的样例标记为正例,而把其它类的样例标记为反例,使用顺序覆盖算法产生区分正例和反例的规则.接下来,Ripper提取区分C_2和其它类的规则,重复该过程,直到剩下类C_k,此时C_k作为默认类.因此最少出现的类最先处理,最常见的类最后处理.Ripper算法使用从一般到特殊的策略进行规则增长,使用FOIL信息增益来选择最佳合取项添加到规则前件中.由于Ripper算法产生规则的特殊性,使得它对于不平衡类的数据集的分类性能较C4.5等算法好.本文选择Ripper算法作为基本算法.

2　基于一趟聚类和Ripper的稀有类分类方法

2.1　算法描述

分类算法通常由分类模型的建立和评估两大块构成,而差别主要体现在模型建立上.本文提出一种基于一趟聚类和Ripper的稀有类分类方法,其建模过程描述如下:

第一步:一趟聚类:对数据集D进行一趟聚类,得到聚类结果$C=\{C_1, C_2, \cdots, C_k\}$;

第二步:划分簇:计算每个簇$C_i(1\leqslant i\leqslant k)$在整个数据集中所占的比例$p_i$,将$p_i$小于等于15%的簇标识为少数簇,而将$p_i$高于15%的类标识为多数簇.

第三步:提取分类规则:利用Ripper分类算法对第二步识别出的少数簇进行分类,提取规则集R1,对多数簇进行分类,提取规则集R2;并最终调整规则集R2-R1为分类多数簇的规则集;R=R1∪(R2-R1),R即为整个数据集的分类规则集.

2.2　阈值选择

在一趟聚类算法中,参数r将影响聚类的结果和算法的时间效率.r越小得到的类的个数越多,算

法时间开销越大,聚类后每个簇中的对象越少,这个簇被划分到稀有类的可能性就越大,当 r 大到一定值时只能得到极少的簇甚至一个簇,此时每个簇中的对象数都很多,甚至没有所占比例低于15%的簇,也就识别不出稀有类了.当 $r=0$ 时,每个簇只有一个元素,也就是说 r 太大或太小都不能得到有意义、有用的聚类结果,这直接影响到少数簇的识别和最终的分类效果.本文采用抽样技术来计算阈值 r 的范围,具体描述如下:

(1)在数据集 D 中随机选择 N 对对象;

(2)计算每对对象间的距离;

(3)计算(2)中距离的平均值 EX 和标准差 DX;

(4) 取 r 在 EX 到 EX $-0.5*$ DX 之间.

3 实验结果及分析

在稀有类分类问题中,更应关注稀有目标类的正确分类率.本文选择如下指标评价分类方法:准确率(accuracy),召回率(recall),精度(precision)和 F-measure.为了验证提出的稀有类分类方法的分类性能,选取 UCI 机器学习数据集[10]中的4个作为实验数据集.4个数据集特征如下:①sick 数据集.有3772条记录,每条记录由26个属性描述,记录分为0和1两个类,0类和1类分别有3541和231条记录,分别占总样本比例93.88%和6.12%,1类可看作稀有类.②hypothy 数据集.有3163条记录,26个属性,包含 hypothyroid 和 negative 两个类别,其中 hypothyroid 类和 negative 类分别拥有实例数目151和3012,分别占总样本比例4.77%和95.23%,hypothyroid 类可看作稀有类.③乳腺癌数据集(Wisconsin breast cancer data set).数据集 breast 有483条记录,其中良性(2类)有444条记录,恶性(4类)有39条记录,每条记录由9个数值属性描述.④KDDCUP99 数据集包含了约4 900 000条模拟攻击记录,总共22种攻击,分为 DOS,R2L,U2R,Probing 等4类;由7个分类特征和34个数值型特征刻画.整个数据集太大,从中随机选取19800条记录的子集,其中攻击记录占257记录.

采用十折交叉验证的方法实验,并将实验结果与相关的分类算法进行对比.实验所用到的 Ripper 分类算法是 Weka[11-12]平台提供的 JRip.表1~表4给出了4个数据集上的对比实验结果.

表1 在 sick 数据集上的稀有类(1类)分类效果比较

算法	Recall	Precision	F-measure	Accuracy
CEEP[7]	67.53	86.67	75.91	97.38
CREEPTP[7]	72.73	87.96	79.62	97.72
BeEPRC[6]	83.55	82.13	82.83	97.88
BeEPRCF[6]	83.98	83.98	83.98	98.04
Ripper 算法	85.7	85.7	85.7	98.25
本文算法	89.19	89.49	88.73	96.03

表2 在 hypothy 数据集上的稀有类(hypothyroid 类)分类效果比较

算法	Recall	Precision	F-measure	Accuracy
BeEPRC[6]	90.07	87.18	88.60	98.89
BeEPRCF[6]	89.40	86.54	87.95	98.83
Ripper 算法	92.1	90.8	91.4	99.18
本文算法	92.50	91.55	91.76	99.21

表3 在 breast 数据集上的稀有类分类效果比较

算法	Recall	Precision	F-measure	Accuracy
Ripper 算法	97.9	91.8	94.8	96.28
本文算法	97.19	85.11	90.33	92.70

表4 在 KDDCUP99 数据集上的稀有类分类效果比较

算法	稀有类	Recall	Precision	F-measure	Accuracy
Ripper 算法	dos	97.6	99.6	98.6	99.93
	probe	20	100	33.3	99.93
本文算法	dos	97.41	99.26	98.28	99.95
	probe	40	40	40	99.95

从表1~表4可以看出,在评价稀有类分类的这4个评估标准中,基于一趟聚类和 Ripper 的稀有类分类算法均能够取得高质量的分类效果.

为考察聚类阈值 r 对分类结果的影响,表5~表9给出了4个数据集在不同 r 下的分类结果.

对于 sick 数据集,经计算求得 EX $=0.30$,DX $=0.11$.从表5可以很明显的看到,参数 r 在 EX $-0.8*$ DX 到 EX $+2$DX 时,可以达到较稳定的聚类效果,当 r 在0.22到0.5间取值时,稀有类分类的召回率得到了很大提高,但这是以牺牲精度为代价的,当 $r=0.5$ 时,精度的牺牲最小,达到高质量的稀有类分类效果.

对于 hypothy 数据集,经计算求得 EX $=0.23$,DX $=0.14$.从表6可以很明显的看到,参数 r 在 EX $-$ DX 到 EX $+$ DX 时,可以达到较稳定的聚类效果.

表5　在sick上参数 r 的选择对稀有类(1类)分类效果的影响比较

r	Recall	Precision	F-measure	Accuracy
0.22	92.79	33.15	47.01	83.35
0.32	89.53	33.84	48.47	87.89
0.42	88.65	62.49	72.78	92.23
0.5	89.19	89.49	88.73	96.03

表6　在hypothy上参数 r 的选择对稀有类(hypothyroid类)分类效果的影响比较

r	Recall	Precision	F-measure	Accuracy
0.16	93.09	91.34	91.88	99.21
0.22	92.50	89.32	90.51	99.176
0.24	93.99	90.39	91.91	99.23
0.32	92.50	87.89	89.79	99.02
0.4	91.79	86.86	88.88	98.80
0.42	91.32	88.38	89.44	98.96
0.48	92.49	92.14	92.08	99.24
0.5	92.50	91.55	91.76	99.21
0.52	92.50	92.14	92.07	99.23

对于breast数据集,经计算求得EX=0.46, DX=0.36.从表7可以很明显地看到,参数 r 在EX-DX到EX时,可以达到较稳定的聚类效果.

表7　在breast上参数 r 的选择对稀有类(4类)分类效果的影响比较

r	Recall	Precision	F-measure	Accuracy
0.1	97.07	91.19	93.80	95.85
0.2	97.19	85.11	90.33	92.70
0.3	96.19	84.97	89.26	90.28
0.4	87.32	78.34	79.21	80.12
0.45	92.17	91.41	91.35	94.27

对于KDDCUP99数据集,经计算求得EX=0.24,DX=0.13.从表8和表9可以很明显的看到,参数 r 在EX-DX到EX+DX时,可以达到较稳定的聚类效果.

表8　在KDDCUP99上参数 r 的选择对稀有类(dos类)分类效果的影响比较

r	Recall	Precision	F-measure	Accuracy
0.1	98.56	98.89	98.68	99.96
0.12	97.27	99.58	98.34	99.96
0.18	97.41	99.26	98.28	99.95
0.24	97.41	90.45	93.29	99.80
0.3	97.71	94.51	95.92	99.87
0.35	98.95	97.58	98.13	99.92
0.4	98.40	23.98	37.4	93.58

表9　在KDDCUP99数据集上参数 r 的选择对稀有类(probe类)分类效果的影响比较

r	recall	precision	F_measure	accuracy
0.1	20	15	16.67	99.96
0.12	40	40	40	99.96
0.18	40	40	40	99.95
0.24	40	40	40	99.80
0.3	40	40	40	99.87
0.35	0	0	0	99.92
0.4	25	30	26.67	93.58

上述实验结果表明,当参数 r 在合适的范围内变化时,尽管聚类结果有变化,出现了聚类合并或分割的现象,以及少量对象由一个簇转移到了另一个簇,但稀有类的识别基本稳定,没有很明显的变化.只是在这个过程中,可能会把多数类标识为稀有类,从而对后面要进行的Ripper分类产生一定的影响,这主要是由于稀有数据在整个数据集中仅占很少的比例,且偏离整个数据集较远,r 的变化主要影响高密度的大簇,而对低密度的小簇影响不大.由于最终目的在于按照15%的比例标识出少数类,而不是中间聚类结果的质量,因此,有理由相信提出的稀有类分类方法对参数 r 分类效果是稳健的.

本文提出了基于一趟聚类和Ripper的稀有类分类方法,先进行一趟聚类,通过将在整个数据集中所占的比例低于15%的聚类标识为少数簇,再应用Ripper分类算法分别对少数簇和多数簇进行分类建模,并按照一定的组合方式调整得出整个数据集的最终规则集.在UCI机器学习数据集上的实验表明,基于一趟聚类和Ripper的稀有类分类方法具有更高的召回率,精度,F-度量值和分类准确率,有利于提高稀有类的分类性能.

[参考文献]

[1] TAEHO J, NATHALIE J. Class imbalances versus small disjuncts[J]. Sigkdd Explorations,2004,6(1):44-49.

[2] ALBERT O, ESTER B M. The class imbalance problem in learning classifier systems: A preliminary study[J]. Genetic And Evolutionary Computation Conference, 2005: 74-78.

[3] GARY M W. Mining with rarity: A uinfying framework [J]. Sigkdd Explorations,2004,6(1): 7-19.

[4] WU Jun-jie, PENG Hui-xiong, CHEN Wu-jian. Local decomposition for rare class analysis[J]. Conference on Knowledge Discovery in Data,2007: 814 - 823.

[5] HAMAD Alhammady, KOTAGIRI Ramamohanarao. The application of emerging patterns for improving the quality of Rare-class classification[J]. Advances in Knowledge Discovery and Data Mining(PAKDD2004),2004: 207 - 211.

[6] 刘艳霞. 基于eEP的稀有类分类问题研究[D]. 郑州: 郑州大学,2005.

[7] 范 明,职为梅. 利用基本显露模式两阶段分类稀有类[J]. 微机发展,2005,15(12):44 - 47.

[8] JIANG Sheng-yi, SONG Xiao-yu. A clustering-based method for unsupervised intrusion detections[J]. Pattern Recognition Letters,2006,5: 802 - 810.

[9] TAN Pang-ning, MICHAEL STEINBACH VIPIN KUMAR 著. 数据挖掘导论[M]. 范明,范宏建等译. 北京: 人民邮电出版社, 2006.

[10] ASUNCION A, NEWMAN D J. UCI Machine Learning Repository[EB/OL] [http://www.ics.uci.edu/~mlearn/MLRepository.html],2007.

[11] WITTEN I H, EIBE FRANK. Data Mining: Practical machine learning tools and techniques[M]. 2nd Edition. San Francisco: Morgan Kaufmann, 2005.

[12] IAN H. Witten and eibe frank. Weka[EB/OL]. http://www.cs.waikato.ac.nz/ml/weka/

[责任编辑:黄建军]

《暨南大学学报(自然科学与医学版)》
再次被收录为"中国科技论文统计源期刊"
再次入编《中文核心期刊要目总览》

"中国科技论文统计源期刊(中国科技核心期刊)"是中国科学技术信息研究所经过严格的定量和定性分析选取的各个学科的重要科技期刊。中国科技论文统计源期刊的论文构成了中国科技论文与引文数据库,该数据库的统计结果编入国家统计局和国家科学技术部编制的《中国科技统计年鉴》,统计结果被科技管理部门和学术界广泛采用。据中国科学技术信息研究所统计,2008年版收录中国科技论文统计源期刊共1 765种。《暨南大学学报(自然科学与医学版)》2008年再次被收录为"中国科技论文统计源期刊"。

《中文核心期刊要目总览》是北京大学图书馆编委会依据文献计量学的原理和方法,采用定量评价和定性评价相结合的方法,从我国正在出版的中文期刊中评选出了1 980余种核心期刊。定量评价指标体系包括被索量、被摘量、被引量、他引量、被摘率、影响因子、获国家奖或被国内重要检索工具收录、基金论文比、Web下载量等9个评价指标;定性评价则采用5 500位学科专家定性评审。《暨南大学学报(自然科学与医学版)》入编《中文核心期刊要目总览》2008年版综合科学技术类的核心期刊,这是继2004年版《中文核心期刊要目总览》之后再次入选。

暨南大学学报编辑部通过不断努力,全面提升了期刊的竞争力和影响力。《暨南大学学报(自然科学与医学版)》再次被收录为"中国科技论文统计源期刊"和再次入编《中文核心期刊要目总览》,是对学报编辑部过去工作的肯定,今后我们将进一步发扬忠信笃敬、知行合一、自强不息、和而不同的暨南精神,把暨南大学学报办成更有影响力的知名学术期刊。

(暨南大学学报编辑部)