非平衡数据集分类问题研究进展*)

高嘉伟[1,2] 梁吉业[1,2]

(山西大学计算智能与中文信息处理省部共建教育部重点实验室 太原 030006)[1]

(山西大学计算机与信息技术学院 太原 030006)[2]

摘 要 非平衡数据集广泛存在于现实世界中,其分类问题已经成为目前数据挖掘领域中的一个研究热点。文章综述了非平衡数据集分类问题的评价方法及其常用分类算法,分析了目前存在的主要困难,并指出需进一步解决的几个问题。

关键词 非平衡数据集,分类,算法

Research and Advancement of Classification Method of Imbalanced Data Sets

GAO Jia-Wei[1,2] LIANG Ji-Ye[1,2]

(Key Laboratory of Ministry of Education for Computation Intelligence & Chinese Information Processing, Shanxi University, Taiyuan 030006)[1]

(School of Computer & Information Technology, Shanxi University, Taiyuan 030006)[2]

Abstract As the imbalanced data sets are widely used in the world, how to classify them has become a hot topic in the research field of data mining. The thesis summarizes the usual evaluation method and the algorithms which are used to classify the imbalanced data sets at present. Main difficulties and the questions which will be solved in the future are also shown in this paper.

Keywords Imbalanced data sets, Classification, Algorithm

1 概述

所谓非平衡数据集,是指同一个数据集中某些类的样本数远大于其他类的样本数,其中样本少的类为少数类(以下称为正类),样本多的类为多数类(以下称为负类)。非平衡数据集中正类样本与负类样本个数不相等,甚至相差悬殊。利用传统的机器学习方法分类,对于正类来说分类准确率很低,而对于负类则相对较高。因为在传统的机器学习方法中,人们往往关注的是负类样本,所以分类器对其分类效果要明显好于正类。非平衡数据在医疗诊断[1,2]、雷达图像监测[3]、诈骗检测[4,5]、金融贷款管理、企业破产预测、电信设备故障预测等领域中有着广泛的应用前景和现实意义。因此如何将非平衡数据集进行正确分类已经成为目前机器学习和数据挖掘的一个研究热点。因为多类问题通常可以简化为两类问题来解决,所以非平衡数据集的分类问题的研究重点也就转化为提高两类问题中正类的分类性能。

2 非平衡数据集分类器性能的评价方法

由于分类错误率只描述了特定判决阈值时的分类器性能,因此不再适合非平衡数据集情况。这就需要一些新的评价方法和指标来描述或者比较非平衡数据集的分类性能。

对于非平衡数据集的每一个测试样本,两类分类器有四种可能的判决结果,以下记:

TP(True Positive):本属于正类且被判别为正类的样本个数;

FP(False Positive):本属于负类且被判别为正类的样本个数;

FN(False Negative):本属于正类且被判别为负类的样本个数;

TN(True Negative):本属于负类且被判别为负类的样本个数。

表1给出了两类数据集的混合矩阵,它是机器学习与模式识别领域中评价分类性能的常用方法[6,7]。为便于说明,设测试集中的正类样本和负类样本总数分别为 N_1 和 N_2,显然有 $TP+FN=N_1$,$FP+TN=N_2$。

表1 两类数据集的混合矩阵

	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

2.1 Accuracy 方法

精确度是分类问题中常用的评价准则[6,7],它反映分类器对于数据集的整体分类性能。其计算公式是 $Accuracy=(TP+TN)/(TP+FP+TN+FN)=(TP+TN)/(N_1+N_2)$。它的使用前提是:类分布是已知且不变的,正类和反类的误分类损失是相等的。但是在非平衡数据挖掘领域,通常不能保证此前提成立。高度的类非平衡通常导致高度非均衡的误分类损失,即正类样本具有较高的误分类损失,误分类一个正类样本的损失远大于误分类一个反类样本的损失。在此

*)本文得到国家自然科学基金(No. 70471003)、高等学校博士学科点专项科研基金(No. 200501080604)、教育部科学技术研究重点项目(No. 206017)和山西省重点实验室开放基金(No. 200603023)的资助。**高嘉伟** 助教,项士研究生,主要研究领域:数据挖掘、智能决策;**梁吉业** 教授,博士生导师,主要研究领域:粗糙集理论、数据挖掘、人工智能等。

情况下，它单独作为评价分类器性能的指标是不够的，它倾向于把样本预测为反类，即反类样本得到较高的分类和预测精度，而我们关注的正类的分类和预测精度却较低。如果将所有的样本都分为反类，精确度仍然很高，但是正类的识别率却为零。因此它并不能单独用来评价非平衡数据集的分类问题。

2.2 F-measure 方法

F-measure 是非平衡数据集分类问题中有效的评价准则[7]，$F\text{-}measure=\sqrt{Recall\times Precision}$，它是查全率(Recall)和查准率(Precision)的组合。

计算正类的 F-measure 时，

$$\begin{cases} Recall=TP/(TP+FN) \\ Precision=TP/(TP+FP) \end{cases}$$

计算反类的 F-measure 时，

$$\begin{cases} Recall=TN/(TN+FP) \\ Precision=TN/(TN+FN) \end{cases}$$

从公式可以看出，如果查全率和查准率的值都比较低，*F-measure* 的值会很小；而如果查全率较高，而查准率较低，同样 *F-measure* 的值也会很小；只有在查全率和查准率的值都比较高的情况下，*F-measure* 的值才会比较大。此方法主要要求在查全率和查准率平衡的前提下尽可能将其最大化。所以 *F-measure* 可以正确地分别评价分类器对于正、负两类的分类性能。

2.3 ROC 曲线

ROC (Receiver Operating Characteristic)曲线能够比较全面地描述分类器在不同判决阈值时的性能[6~8]，所以其成为目前评价非平衡数据集分类器性能的主流方法。

该方法定义了正类准确率 $TPR=TP/(TP+FN)=TP/N_1$，负类准确率 $FPR=FP/(FP+TN)=FP/N_2$。它以 *FPR* 和 *TPR* 分别作为横纵坐标，每一个阈值分别对应一个点(*FPR*，*TPR*)，不断改变阈值就把所有的(*FPR*，*TPR*)点连接起来，这也就是分类器在该测试集上的 ROC 曲线，如图 1 所示。显然，ROC 曲线越靠近左上角表示分类器性能越好。

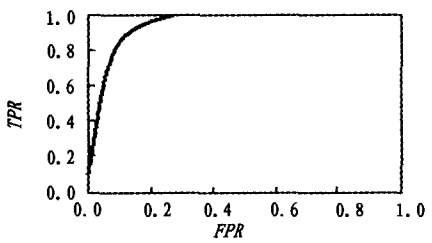

图 1 ROC 曲线

虽然 ROC 曲线比较全面地描述了分类器的分类性能，但是作为一种二维图形描述，ROC 曲线不能给出分类器性能的定量评价。因此，人们常常采用 ROC 曲线下的面积 AUC (the Area Under the Curve)来代替 ROC 曲线本身对分类器的性能进行评估。显然，AUC 取值范围在 0 和 1 之间，并且 AUC 越大，分类器的性能就越好。

3 非平衡数据集分类问题的常用方法

非平衡数据集通常是指两类问题中的负类样本个数远大于正类样本个数，而且正类样本往往是分类问题的关注所在。针对非平衡数据集的分类问题，现有的研究主要包括三方面内容：

(1)利用实验研究类分布对各种传统分类算法的影响[9]。

(2)采用适当的方法重构训练样本数据集[10,11]。

(3)直接改进传统分类算法来提高正类的分类性能[2,8,12,13,15~20]。许多分类方法的设计是基于数据平衡分布假设的，比如线性判别分析方法[8]、Boosting 方法[12,13]以及支撑向量机方法[2,15~18]等。当把这些分类方法应用于非平衡数据集时，就会导致训练出的分类器性能下降。因此，目前人们处理非平衡数据集的分类问题主要是在传统的机器学习方法上做进一步的改进，使之适用于非平衡数据集。这样一方面充分利用现有的信息，另一方面又基本上不增加算法的计算复杂度。

3.1 随机抽样方法

随机抽样方法是直接对训练集进行预处理，然后用处理过的数据集训练分类器，其中有上抽样(up-sampling)和下抽样(down-sampling)之分[10]。上抽样方法通过增加正类样本来提高正类的分类性能，最简单的上抽样方法是复制正类样本，缺点是没有给正类增加任何新的信息，会使分类器学到的决策域变小，从而导致过学习；下抽样方法通过减少负类样本来提高正类的分类性能，最简单的下抽样方法是随机地去掉某些负类样本来减小负类的规模，其缺点是会丢失负类的一些重要信息[11]。因此，人们提出了许多改进的随机抽样方法。例如，在正类中加入随机高斯噪声，或者产生新的合成正类样本等方法，这样可以在一定程度上避免随机上抽样中出现的过学习问题；在负类中去掉远离分类边界或者引起数据重叠的样本，得到的分类效果会比随机下抽样理想一些。

3.2 Fisher 线形判别分析改进算法

Fisher 线形判别算法(Fisher Linear Discriminant，简称 FLD)具有计算简单、在一定条件下能够实现最优分类的性质，因此成为一种实际应用非常广泛的分类方法。但是当两类样本的协方差矩阵不同时，样本不平衡会导致 FLD 的性能下降。因此在此基础上，文[8]提出了一种加权 Fisher 线性判别方法 WFLD (Weighted Fisher Linear Discriminant)以减小样本非平衡带来的影响。WFLD 模型能有效地提高 FLD 在非平衡数据集上的泛化性能。但究其实质，WFLD 实质上等价于一种特殊的上抽样方法，即同时对两类样本进行不同倍数的上抽样，使两类样本个数之比为 1:1，也就是相当于使原始非平衡数据集变为两类样本个数为 1:1 的平衡数据集。此方法虽然从表面上看消除了样本非平衡对分类性能的影响，但是并没有实质解决上抽样方法的固有缺点。

3.3 Boosting 改进算法

传统的 Boosting 算法是一种提高任意给定学习算法准确度的方法，其核心是一种迭代算法[12,13]。初始化时给训练集中每个训练样本的权重都指定相同的分布。然后进行迭代，每次迭代后，按照训练结果更新训练集上的分布，对于训练失败的训练样本赋予较大的权重，使得下一次迭代更加关注这些训练样本。Boosting 算法具有很多优点，它有较高的正确率，不需要先验知识，只需要选择合适的迭代次数等。

文[12]在 AdaBoost 算法的基础上提出了 RareBoost-Ⅰ算法，其核心是正类和负类迭代的权重值可以取不同值。而在 Boosting 家族中 SLIPPER 算法的基础上改进的 RareBoost-Ⅱ算法，在每次迭代时只将某一类重建模型改进为在每次迭代时正、反类都要重新构建模型。这两种方法都要直接影响正类的两个参数：查全率 Recall(即 $TP/(TP+FN)$)

和查准率 Precision(即 $TP/(TP+FP)$),然后将数据重新平衡。但是文中没有对 TP,FP,TN 和 FN 这四个不同因子在迭代中权重值更新的规律进一步说明和论证。

文[13]提出了 DataBoost-IM 方法,此方法将正类与反类中的代表样本选中,并添加到原数据集中,直到新训练集中权重值和类分布重新平衡,但是如何寻找两类中代表性的样本是一个难点。

这两种方法主要是采用通过调整权重值使得数据重新平衡,在此基础上再利用传统机器学习方法进行分类。但是 Boosting 改进算法并没有改变 Boosting 算法的原有缺点,即如何选择合适的迭代次数,如何减少噪声的影响,因为它在迭代过程中总是给噪声分配较大的权重,使得这些噪声在其后的迭代中受到更多的关注。

3.4 支撑向量机改进算法

支撑向量机(Support Vector Machine, SVM)是一种强大的机器学习和数据挖掘算法。由于其坚实的理论基础和良好的泛化性能,SVM 已经成为最流行的分类算法之一[14]。

文[15]分别从理论和仿真实验的角度论证了类分布对于 SVM 分类的影响,确定了支持向量 SV 数(率)和边界支持向量 BSV 数(率)的界,并分别推广到正类和反类。它从理论上分析了基于精度的分类算法对于非平衡数据挖掘为何表现出对正类的分类和预测较差,而反类的分类和预测精度高的原因。这对于加权 SVM 补偿类分布以及改进其它基于精度的分类算法提供了理论指导。

文[16]提出了样本数目不对称时的 SVM 模型,该模型关键在于如何确定 $C_i(i=1,2)$值,其中常数 C_i 起着对错分样本的惩罚作用,实现的是学习机器泛化能力和错分样本数目之间的折中。C_i 值太小,表明对样本的错分惩罚太小,可能会导致过多的错分样本;C 值太大,表明样本的错分惩罚大,同样起不到折中的作用。

文[2]利用 BMPM(Biased Minimax Probability Machine)算法来解决 FP (Fractional Programming)最优化的问题,它通过训练集估计出分类的均方差和协方差矩阵,再利用这两个矩阵推算未来样本集误分类的最坏情况的边界,它不再假定 MPM 中每类的无偏权重,而直接用量化的方法来控制超平面向正类移动。但是这种方法要求估计的均方差和协方差矩阵要尽可能准确。

文[17]利用 ACT (Adaptive Conformal Transformation)算法改变了特性空间距离和类不平衡率,以提高支撑向量机在非平衡数据集中决定类边界的性能,但是算法中参数 BM (Boundary Movement)、BP(Biased Penalties)和 KM(Kernel Modification)的选择和其他邻近的支持向量机的选择较为困难。文[18]基于 SVM 提出了 KBA(Kernel Boundary Alignment)算法,它由训练集得到先验知识来提高类预测的准确性,但是如何由非平衡数据集得到先验知识,并由其估算出"理想"边界较为困难。

3.5 基于信息粒的分类方法

文[6]从信息粒的角度提出了 KAIG (Knowledge Acquisition via Information Granulation)模型,并利用同质指数 H-index 和不可分辨比率 U-ratio 两个参数来构造合适的粒度,并定义了次属性来描述和解决粒重叠的问题,取得了一定的突破。这个模型不仅可以对平衡数据集进行分类,而且可以突出那些重要的数据,这对于解决非平衡数据集的分类问题提供了有效的方法。但是文章仅对定量属性进行说明,而对于具有定性属性或者混合属性(即同时具有定性属性与定量属性)的非平衡数据集的分类问题并没有做进一步的研究。

3.6 其他算法

此外一些学者利用决策树[19,20]、粗糙集和模糊集[1]等方法探讨了非平衡数据集的分类问题。

4 非平衡数据集分类问题中进一步研究的几个问题

虽然国内外学者在非平衡数据集分类问题的研究方面已经取得了一些成绩,但这仅仅是初步的研究,还面临一些亟需解决的问题。

4.1 如何减少分类过程中正类信息的丢失或者失真

正类的信息是非平衡数据集中最有用的信息,也是主要研究的对象,很多分类方法都是通过直接"复制"或者系数调整参数修正的方法使正类放大,所以在分类过程中如何才能尽可能防止其丢失或者失真是我们需要面对和解决的问题。

4.2 如何由非平衡数据集的先验知识得到能被训练集使用的较为准确的参数

从前面列举的一些改进算法可以看出,很多是由先验知识得到训练集的参数,并由这些参数将非平衡数据集重新调整平衡。因此如何较准确地获得这些参数是一个非常关键的问题。

4.3 如何对具有定性属性或者混合属性的非平衡数据集进行正确分类

尽管人们利用信息粒的思想构建的 KAIG 模型成功地解决了只具有定量属性的非平衡数据集的分类问题,但是在非平衡数据集中有相当一部分信息是具有定性属性或者混合属性的。因此如何解决这个问题就成为目前研究的一个新的热点和难点。

结束语 非平衡数据集的分类问题是当前机器学习和模式识别领域中新的研究热点之一,是对传统分类方法的重大挑战。虽然目前国内外学者对这一问题也展开了初步的研究和讨论并取得了一些成果,但是仍然有诸多问题亟需解决。我们认为非平衡数据集的分类问题在今后很长一段时间内仍将是一个研究热点,新的成果也将会不断涌现,从而推动非平衡数据集分类技术更快地发展和更广泛地应用。

参 考 文 献

1 Wilk S, Slowinski R, Michalowski W, et al. Supporting triage of children with abdominal pain in the emergency room. European Journal of Operational Research, 2005, 160(3): 696～709

2 Huang K Z, Yang H Q, Irwin K, et al. Biased Mimimax Probability Machine for Medical Diagnosis. In: Proc. of the 8th International Symposium on Artificial Intelligence and Mathematics, Florida, 2004

3 Miroslav K, Robert C H, Stan M. Machine learning for the detection of oil spills in the satellite radar images. Machine Learning, 1998, 30: 195～215

4 Tom F, Foster P. Adaptive fraud detection. Data mining and knowledge discovery, 1997, 3(1): 291～316

5 Chan P K, Stolfo S J. Toward Scalable Learning with Non-UniformClass and Cost Distributions: A Case Study in Credit Card Fraud Detection. In: Proc. of the 4th International Conference on Knowledge Discovery and Data Mining, New York, 1998. 164～168

6 Su C T, Chen L S, Yih Y. Knowledge acquisition through information granulation for imbalanced data. Expert Systems with applications, 2006, 31: 531～541

7 Joshi M V. On Evaluating Performance of Classifiers for Rare Classes. In: Proc. of the 2nd IEEE International Conference on Data Mining, Maebishi, Japan, 2002. 641～644

8 谢纪刚，裘正定. 非平衡数据集 Fisher 线性判别模型. 北京交通大学学报，2006，30(5)：15～18
9 Weiss G. Mining with Rarity Problems and Solutions: A Unifying framework. SIGKDD Explorations, 2004, 6(1): 7～19
10 Marcus A. Learning when data sets are imbalanced and when costs are unequal and unknown. In: Proc. of the Workshop on Learning from Imbalanced Data Sets II, ICML, Washington DC, 2003
11 Miroslav K, Stan M. Addressing the curse of imbalanced training sets-one sided selection. In: Proc. of the 14th International Conference on Machine Learning, Morgan Kaufmann, 1997. 179～186
12 Mahesh V, Vipin K, Ramesh C. Evaluating boosting algorithms to classify rare classes: comparison and improvements. In: Proc. of the 1st IEEE International Conference on Data Mining, San Jose, CA, 2001. 257～264
13 Guo H Y, Herna L V. Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach. Sigkdd Explorations, 2003, 6: 30～39
14 Vapnik V. The nature of statistical learning theory. New York: Springer Verlag Press, 1995
15 郑恩辉，李平，宋执环. 不平衡数据知识挖掘：类分布对支持向量机分类的影响. 信息与控制，2005，34(6)：703～708
16 肖健华，吴今培. 样本数目不对称时的 SVM 模型. 计算机科学，2003，30(2)：165～167
17 Wu G, Edward Y C. Adaptive Feature-Space Conformal Transfor-mation for Imbalanced-Data Learning. In: Proc. of the 20th international conference on machine learning, Washington DC, 2003. 816～823
18 Wu G, Edward Y C. KBA: Kernel boundary alignment considering imbalanced data distribution. The IEEE transactions on knowledge and data engineering, 2005', 17(6): 786～795
19 Claire C, Nicholas H. Improving minority class prediction using case-specific feature weights. In: Proc. of the 14th international conference on machine learning, Morgan, Kaufmann, 1997. 57～65
20 Chris D, Robert C H. C4. 5, class imbalance, and cost Sensitivity: Why under-sampling beats over-sampling. In: Proc. of the Workshop on Learning from Imbalanced Data Sets II, ICML, Washington DC, 2003

（上接第 9 页）

为了解决以上技术问题，研究人员在多个方面探索解决流量矩阵估计问题的最佳方案。测量反演结合策略，是未来继续发展的一个方向。由于随着网络的发展，目前有许多的流监测器对我们而言是可以直接使用的，从而使得这一方法可行，并成为未来的一个发展方向。目前获得流量矩阵的已知方法大多仅仅依靠 SNMP 数据来进行，这些方法对较小流量的处理是很困难的，测量反演结合策略的下一代方法将能够更好地处理这一问题[13]。

充分考虑 OD 流的时间、空间或者是时间空间相关性，建立合理的 OD 流模型，更好利用测量附加信息，提高估计精度，降低误差，是以后流量矩阵估计的进一步发展方向。对 OD 流动态变化自适应性的提高，也是以后我们需要解决的问题。

参 考 文 献

1 Cao J, Davis D, Weil S V, et al. Time-varying Network Tomography. J of the American Statistical Association, 2000
2 Vardi Y. Network tomography: Estimating source-destination traffic intensitiesfrom link data. Amer J Stat Assoc, 1996, 91 (433): 365～377
3 Tebaldi C, West M. Bayesian inference on network traffic using link count data. Journal of the American Stati-stical Association, 1998, 93(442): 557～576
4 Tsang Y, Coates M, Nowak R. Passive network tomography using EM algorithms. In: Proc. IEEE Int Conf. Acoust, Speech, and Signal Proc., May 2001
5 Coates M, Hero A, Nowak R, et al. Internet tomography. IEEE SignalProcess, 2002, 19(3): 47～65
6 Tsang Y, Coates M, Nowak R. Nonparametric internet tomography. In: Proc. IEEE Int Conf. Acoust. Speech, Signal Process, May 2002
7 Chen, Bindel K. Tomography-based Overlay Network Monitoring. ACM IMC, 2003
8 Duffield N. Simple Network Performance Tomography. ACM IMC, 2003
9 Zhang Y, Roughan M, Duffield N, et al. Fast Accurate Computation of Large-Scale IP Traffic Matrices from Link Loads. In: ACM Sigmetrics, San Diego, CA, 2003
10 Zhang Y, Roughan M, Lund C, et al. Estimating Point-to-Point and Point-to-Multipoint Traffic Matrices: An Information-Theoretic Approach. IEEE/ACM Transactions on Networking, 2004
11 Zhang Y, Roughan M, Lund C, et al. An Information Theoretic Approach to Traffic Matrix Estimation. In: ACM SIGCOMM, Karlsruhe, Germany, August 2003
12 Medina, et al. An Information-Theoretic Approach to Traffic Matrix Estimation. ACM SigComm, 2002
13 Soule A, Lakhina A, Taft N, et al. Traffic Matrices: Balancing Measurements, Inference and Modeling. ACM Sigmetrics 2005. Banff, June 2005
14 Medina A, Taft N, Salamatian K, et al. Traffic Matrix Estimation: Existing Techniques and New Directions. In: ACM SIGCOMM, Pittsburgh, USA, Aug. 2002
15 Soule A, Salamatian K, Nucci A, et al. Traffic Matrix Tracking using Kalman Filtering: [Research Report]. RP-LIP6-2004-07－10. LIP6, 2004
16 张宏莉，方滨兴，胡铭曾，等. Internet 测量与分析综述. 软件学报，14(1)：110～116
17 Kay S M 著. 统计信号处理基础：估计与检测理论. 电子工业出版社，2003
18 张贤达. 矩阵分析与应用. 清华大学出版社，2004
19 Bhattacharyya S, Diot C, Jetcheva J, et al. Geographical and Temporal Characteristics of Inter-POP Flows: View from a Single POP. In: European Transactions on Telecommunications, February 2002
20 Nucci A, Cruz R, Taft N, et al. Design of IGP Link Weight Changes for Estimation of Traffic Matrices. In: IEEE Infocom, Hong Kong, March 2004
21 Soule A, Nucci A, Leonardi E, et al. How to Identify and Estimate the Largest Traffic Matrix Elements in a Dynamic Environment. In: ACM Sigmetrics, New York, June 2004
22 Gunnar A, Johansson M, Telkamp T. Traffic Matrix Estimation on a Large IP Backbone - A Comparison on Real Data. In: ACM Internet Measurement Conference, Taormina, Italy, October 2004
23 Papagiannaki K, Taft N, Lakhina A. A Distributed Approach to Measure Traffic Matrices. In: ACM Internet Measurement Conference, Taormina, Italy, October 2004
24 Erlander S, Stewart N F. The Gravity Model in Transportation Analysis - Theory and Applications. 1990
25 Vanderbei R J, Iannone J. An EM approach to OD matrix estimation: [Tech Rep]. SOR 94-04, Princeton University, Princeton, NJ, 1994
26 Lakhina A, Papagiannaki K, Crovella M, et al. Structural Analysis of Network Traffic Flows. In: ACM Sigmetrics, New York, June 2004
27 Rschendorf L. Convergence of the iterative proportional fitting procedure. Annals of Statistics, 1995. 1160～1174
28 Liang G, Taft N, Yu Bin. A Fast Lightweight Approach to Origin-Destination IP Traffic Estimation Using Partial Measurements. IEEE Transactions on Informationtheory, 2006, 52(6)