不平衡数据集的分类方法研究

王和勇[1]，樊泓坤[2]，姚正安[2]，李成安[1]
(1.华南理工大学 电子商务学院，广州 510006；2.中山大学 数学与计算科学学院，广州 510275)

摘 要：传统的分类算法在处理不平衡数据分类问题时会倾向于多数类，而导致少数类的分类精度较低。针对不平衡数据的分类，首先介绍了现有不平衡数据分类的性能评价；然后介绍了现有常用的基于数据采样的方法及现有的分类方法；最后介绍了基于数据采样和分类方法结合的综合方法。

关键词：机器学习；不平衡数据；数据分类
中图分类号：TP181 **文献标志码**：A **文章编号**：1001-3695(2008)05-1301-03

Research of imbalanced data classification

WANG He-yong[1], FAN Hong-kun[2], YAO Zheng-an[2], LI Cheng-an[1]
(1. *College of E-business, South China University of Technology, Guangzhou* 510006, *China*; 2. *College of Mathematics & Computer Science, Sun Yat-sen University, Guangzhou* 510275, *China*)

Abstract: Imbalanced data set cause the deduction of the precision of the minority class samples, when it is classified by traditional algorithm, which can tend to favor the more class samples. In view of the imbalanced data classification, this paper firstly introduced the developed methods that were the performance evaluation of imbalanced data classification. Secondly it presented the developed sampling methods regarding imbalanced data set and produced the classified methods. In the end, it showed the union methods of using sampling method and classified method.

Key words: machine learning; imbalanced data set; data classification

在过去的几十年中，全球信息科技的飞速发展导致了功能强大的计算机、数据收集设备和存储设备的产生。利用这些设备可以收集大量的数据信息以供人们进行事务管理、信息检索和数据分析。尽管收集得到的数据量非常大，但是对人们有用的数据往往非常有限，通常仅占全部数据的一小部分。这种某类样本数量明显少于其他类样本数量的数据集称为不平衡数据集。不平衡数据集的分类问题大量存在于人们的现实生活和工业生产之中。例如，寻找电信运行商的逃离客户[1]，一般情况下逃离的客户要远远少于非逃离客户；利用检测数据诊断病人的疾病[2]，如癌症，人们患癌症的概率是非常低的，因此癌症患者要远远少于健康的人；其他如从卫星图片中油井的定位[3]、学习单词的发音[4]、文本自动分类[5]、分辨恶意的骚扰电话[6]等。在这些应用中，人们主要关心的是数据集中的少数类，而且这些少数类的错分所产生的代价非常大。把有逃离倾向的客户判为正常客户将有可能失去该客户；把癌症病人误诊为正常将会延误治疗时机，对病人造成生命威胁。因此在实际应用中，需要提高少数类的分类精度。

近几年来，不平衡数据集的分类问题也越来越受到数据挖掘和机器学习学术界的重视，已成为数据挖掘和机器学习界的热点问题之一。2000 年美国人工智能协会(AAAI)[7]以及 2003 年机器学习国际会议(ICML)[8]特别对不平衡数据的学习问题召开了专题讨论会。2004 年美国计算机协会(ACM)针对这一专题出版了一期通讯[9]。目前，处理不平衡数据集分类较好的方法主要有基于不平衡数据集的数据采样和基于不平衡数据集的分类。

1 不平衡数据分类的性能评价

表 1 是二类问题的混淆矩阵。表中 *TP* 是真实正例的数目；*FP* 是虚假正例的数目；*TN* 是真实负例的数目；*FN* 是虚假负例的数目。

表 1 二类问题的混淆矩阵

真实类标	预测类标		
	正例	负例	Σ
正例	*TP*	*FN*	n^+
负例	*FP*	*TN*	n^-

在一般情况下，常用精确率：accuracy = $(TP + TN)/(n^+ + n^-)$ 来评价分类器的性能。但在处理不平衡数据集的分类问题时，这种评价方法显然不合适。例如数据集不平衡，仅有 5%的正例，那么分类器只需把所有测试样本分为负例，即可获得很高的精确率(95%)。但这样的分类器是没有实际意义的。因此，对于不平衡数据的分类应考虑使用其他评价方法。

受试者工作特性(receiver operating characteristic, ROC)曲线[10]以及正负例精确率的几何平均[11]是两种流行的分类器性能评价方法。它们都独立于数据集类间的分布，对数据集的不平衡性有很好的鲁棒性，因此它们可用于不平衡数据集分类器的评价。

ROC 曲线图的纵轴表示真实确定率：

$$TP \text{ rate} = TP/n^+ \quad (1)$$

收稿日期：2007-05-06；修回日期：2007-07-21

作者简介：王和勇(1973-)，男，河南泌阳人，博士，主要研究方向为数据挖掘、商业智能(zsuwhy@hotmail.com)；樊泓坤(1979-)，男，硕士，主要研究方向为数据挖掘；姚正安(1960-)，男，教授，主要研究方向为计算机通信、图像处理、模式识别；李成安(1969-)，男，工程师，博士，主要研究方向为模式识别、智能系统.

横轴表示虚假确定率:

$$FP \text{ rate} = FP / n^- \quad (2)$$

ROC 曲线反映了当分类器参数变换时,真实确定率(*TP* rate) 与虚假确定率(*FP* rate) 之间的关系。坐标中的(0, 0) 点对应于将所有点归于负类; (1, 1) 点对应于将所有点归于正类; 直线 $y = x$ 对应于随机猜测; (0, 1) 点对应于最理想的分类情况,即所有样本均分类正确。ROC 曲线越靠近左上角,分类器性能越好。由于 ROC 曲线没有给出具体的评价数值,不方便不同分类器间性能的比较,于是人们常使用 ROC 曲线下的面积(area under ROC, AUC) 作为评价指标。较大的 AUC 值对应于较优的分类器。AUC 是基于 ROC 曲线的惟一数值,它与错分代价无关,不受与规则应用相关的因素影响。文献[12] 中提出了一种 AUC 值的估计方法,该方法估计 AUC 仅需使用后验概率的排列值,而无须知道具体的 ROC 曲线。

正负例精确度的几何平均计算式如下:

$$g = \sqrt{acc^+ \times acc^-} \quad (3)$$

其中: $acc^+ = TP / n^+$ 为正例的精确率; $acc^- = TN / n^-$ 为负例的精确率。g 值的大小与 ROC 点离最优分类的距离密切相关[13]。用 g 值来衡量分类器处理不平衡数据的性能是直观和合理的。要使 g 值大,则需要正负例的精确率都大,且尽量保持两者平衡; 如果负例的精确率大,而正例的精确率小,那么 g 值仍会比较小。此外,由式(3) 可以看出, g 值与 acc^+、acc^- 的关系是非线性的, acc^+(或 acc^-) 值的变化对 g 值的影响依赖于 acc^+(或 acc^-) 值的大小: acc^+(或 acc^-) 值越小,它引起的 g 值的变化越大。这就意味着少数类即正类样本错分得越多,正类的错分代价就越大。

2 常用的不平衡数据分类方法

处理不平衡数据集分类的方法主要可分为三大类: 基于数据采样、基于分类算法以及将两种类型方法结合的综合方法。基于数据采样的方法主要是改变不平衡数据的分布,以降低数据的不平衡程度; 基于分类算法的方法主要是提出新的分类思想,改进传统的分类算法,以适应不平衡数据分类的需要; 综合方法则是基于数据采样与分类方法的结合。

2.1 基于数据采样的方法

针对原始数据的处理,提出了多种不同的数据重采样方法[14]。按照对样本数量的影响可分为: 向上采样,即人为地增加少数类的样本; 向下采样,即人为地减少少数类的样本。根据方法的智能性又可分为启发式方法和非启发式方法。下面简单介绍一些常用的方法:

a) 随机向上采样。该方法是最简单的向上采样方法,它通过随机复制少数类的样本来达到增加少数类样本的目的。由于这种方法仅仅是对少数样本的简单复制,容易造成分类器的过学习。

b) 随机向下采样。该方法是最简单的向下采样方法,它随机去掉多数类的样本,以降低数据不平衡的程度。由于随机性,这种方法常常会去掉一些潜在的对分类有用的样本。

以上两种方法的原理简单,都属于非启发式方法。下面介绍的方法都属于启发式方法,它们引入了启发式规则,从而可在一定程度上克服非启发式方法的不足。

c) Tomek links。该方法在文献[15] 中提出,其基本思想如下: 给定两个样本 x_i, x_j 属于不同的类,它们之间的距离用 $d(x_i, x_j)$ 表示。若不存在另一样本 x 满足 $d(x_i, x) < d(x_i, x_j)$ 或 $d(x_j, x) < d(x_i, x_j)$,则样本对(x_i, x_j) 构成一个 Tomek links。如果两个样本构成 Tomek links,则其中某个样本为噪点,或者两个样本在两类的边界上。利用这个性质,Tomek links 可作为向下采样的方法,即去掉构成 Tomek links 的负例。

d) 压缩最近邻(condensed nearest neighbor rule, CNN) 。该方法最早由 Hart[16] 提出,当时用于寻找样本的一致子集。子集 $\tilde{E}$ 与 E 称为集合 E 的一致子集。如果用最近邻分类器, $\tilde{E}$ 中样本可完全正确地分类 E 中的样本。文献[11] 提出了一种构造子集 $\tilde{E}$ 的算法作为向下采样的方法: 首先将所有正例样本以及随机选取一个负例加入 $\tilde{E}$ 中进行初始化; 然后用 $\tilde{E}$ 中的样本以最近邻算法(1-NN) 对 E 中样本分类,将所有错分的样本加入到 $\tilde{E}$ 中。这种算法产生的一致子集并不一定是最小的,但它保留了负例边界附加的样本; 同时去掉了负例中远离边界的样本,从而达到减少负例的目的。

e) 邻域清理(neighborhood cleaning rule, NCL) [17]。该方法利用 Wilson 改进的最近邻规则(edited nearest neighbor, ENN) [18] 对多数类进行向下采样。ENN 的基本思想是去掉那些类标与离它最近的三个样本中的两个类标不同的样本。但多数类的样本附近通常都是多数类的样本,因此 ENN 去掉的样本是非常有限的。NCL 对 ENN 作了一点改进,以去掉更多的样本。其基本算法如下: 对训练集中的每个样本 X 找出离它最近的三个样本,若 X 为负例即属于多数类,且三个最近邻样本中有两个以上为正例,则去掉 X; 若 X 为正例即属于少数类,且三个最近邻样本中有两个以上为负例,则去掉三个最近邻样本中的负例。

f) 虚拟少数类向上采样(synthetic minority over-sampling technique, SMOTE) 。该方法是 Chawla 等人[19] 提出的一种向上采样方法。它建立这样的假设之上: 相距较近的正例之间的样本仍是正例。其主要思想是在相距较近的正例之间插入人造的正例。具体算法如下: 对少数类的每一个样本 X,搜索其 k (通常取为 5) 个最近邻; 然后随机选取这 k 个最近邻中的一个设为 $\tilde{X}$,再在 X 与 $\tilde{X}$ 之间进行随机线性插值,构造出新的少数类样本,即新样本

$$x_{new} = x + \text{rand}(0, 1) \times (\tilde{x} - x) \quad (4)$$

其中: rand(0, 1) 表示区间(0, 1) 的一个随机数。若需要更多的虚拟样本,重复以上步骤即可。SMOTE 使分类器的分类平面向多数类的空间伸展,同时可有效地避免随机向上采样的过学习问题。

图 1 是 SMOTE 算法的效果, (a) 为原始样本分布图; (b) 为使用 SMOTE 方法增加了两倍虚拟样本后的分布图。可以看出 SMOTE 方法增加的虚拟样本基本保持了原始样本的分布,但增加的虚拟样本大多分布于原始样本内部,而在边缘附近则较少。

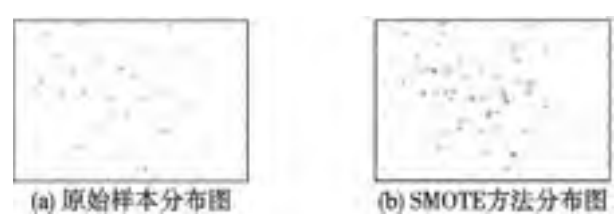

图 1 SMOTE 方法效果图

以上是基于数据采样的几种处理不平衡数据的基本方法。近几年专家们针对这些基本方法的不足提出了许多改进的新方法。例如, Kubat 等人[11] 提出的单边选择(one-sided selec-

tion, OSS) 的向下采样方法, 将 Tomek links 方法与 CNN 方法结合起来; Gustavo 等人将向上采样与向下采样方法结合[14], 提出 SMOTE + Tomek links 和 SMOTE + ENN 方法; Taeho 等人提出基于聚类的向上采样方法[20], 可同时处理类间不平衡以及类内不平衡; Guo Hong-yu 等人在进行 boosting 过程中寻找各类的硬样本, 再利用这些特殊样本生成新的虚拟样本[21]; Han Hui 等人提出 Borderline-SMOTE 方法[22], 仅对少数类的边界数据进行向上采样。

2.2　分类方法

分类方法针对的是分类算法而不是数据集, 它们通过改进现有分类算法来处理不平衡数据集。下面介绍这类方法的一些研究进展。

标准的 Boosting 算法如 Adaboost[23] 没有考虑不平衡数据集的特点, 错分样本所增加的权重与正分样本所减少的权重比例是相同的, 因此传统的 Boosting 算法对少数类的效果不佳。Joshi 等人针对这点不足, 提出了一种改进的 Boosting 算法, 在更新权重时, 赋予预测正例 (*TP*, *FP*) 权重和预测负例权重 (*TN*, *FN*) 不同的改变量。这种算法有效提高了正例预测的精度[24]。

支持向量机(support vector machine, SVM) 在处理不平衡数据时, 分类面会偏向于少数类, 因此会增加少数类的错分率。Wu Gang 等人提出了一种边界调准算法, 通过修改 SVM 的核函数来调整分类面[15]。

Huang Kai-zhu 等人提出 biased minimax probabilitiy machine(BMPM) 方法来解决不平衡的问题[25]。只要知道各类样本可靠的均值和协方差矩阵, BMPM 就可通过调整测试集实际准确率的下界来求出决策超平面。此外, 还有许多有效的分类算法, 如基于代价的学习[26]、one-class 学习[27] 等。

基于分类算法的方法不改变样本的分布, 其基本原理是使分类器更注重少数类, 对少数类的样本更加敏感。但当少数类样本不能反映其真实分布时, 这类算法容易出现过学习现象。

2.3　综合方法

基于数据采样的方法和基于分类算法的方法都有自身的长处和不足, 目前越来越多的研究将两种类型方法结合。结合方法是对原始数据进行重采样, 以降低数据的不平衡性, 然后采用可补偿数据不平衡的分类算法进行分类。例如 Rehan Akbani 等人[28] 所使用的 SMOTE + biased-SVM 方法, 首先使用 SMOTE 方法对少数类向上采样, 降低数据的不平衡程度; 然后在分类算法上使用 Veropoulos 等人[29] 提出的 biased-SVM 方法赋予正负例不同的错分代价。Rehan Akbani 等人通过对比实验验证了他们的方法要优于单纯使用基于数据采样或基于分类算法的方法。

这类算法有效的关键在于如何将两种类型的方法有机结合, 使得既可有效发挥两种类型方法的长处, 同时又可避免各自的弱点。目前两种类型方法结合的综合方法还比较少, 但无论在理论还是实践上, 这类方法都表现出一定的优越性。

3　结束语

不平衡数据集的分类问题是数据分类的难题之一, 其困难主要是由不平衡数据集自身的特点以及传统分类算法的局限性造成的。本文首先介绍了不平衡数据分类的性能评价标准和针对不平衡数据本身的数据采样方法; 然后介绍了针对不平衡数据集的改进分类方法; 最后介绍了利用不平衡数据本身的数据采样与不平衡数据集分类两者结合的方法。

参考文献:

[1] ZAWA K J, SINGH M, NORTON S W. Learning goal oriented Bayesian networks for telecommunications management[C] //Proc of the 13th International Conference on Machine Learning. San Fransisco: Morgan Kaufmann, 1996: 139-147.

[2] CHAWLA N V, BOWYER K W, HALL L O, *et al*. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[3] KUBAT M, HOLTE R, MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning, 1998, 30(2) : 195-215.

[4] BOSCH A T, HERIK H J, DAELEMANS W. When small disjuncts abound, try lazy learning: a case study[C] //Proc of the 7th Belgian-Dutch Conference on Machine Learning. 1997: 109-118.

[5] ZHENG Zhao-hui, WU Xiao-yun, SRIHARI R. Feature selection for text categorization on imbalanced data[J]. SIGKDD Explorations, 2004, 6(1) : 80-89.

[6] FAWCETT T, PROVOST F. Combining data mining and machine learning for effective user profile[C] //Proc of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 8-13.

[7] JAPKOWICZ N. Learning form imbalanced data sets: a comparison of various strategies, WS-00-05[R]. Menlo Park: AAAI Press, 2000.

[8] CHAWLA N V, JAPKOWICZ N, KOLCZ A. Proceedings of the ICML workshop on learning from imbalanced data sets[C]. 2003.

[9] CHAWLA N V, JAPKOWICZ N, KOLCZ A. Editorial: special issue on learning from imbalanced data sets[J]. ACM SIGKDD Exploration Newsletter, 2004, 6(1) : 1-6.

[10] BRADLEY A. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30(6) : 1145-1159.

[11] KUBAT M, MATWIN S. Addressing the course of imbalanced training sets: one-sided selection[C] //Proc of the 14th International Conference on Marchine Learning. San Fransisco: Morgan Kaufmann, 1997: 179-186.

[12] HAND D J, TILL R J. A simple generalisation of the area under the ROC curve for multiple class classification problems[J]. Machine Learning, 2001, 45(2) : 171-186.

[13] BARANDELA R, VALDOVINOS R M, SANCHEZ J S. New applications of ensembles of classifiers[J]. Pattern Analysis and Applications, 2003, 6(3) : 245-256.

[14] GUSTAVO E A, BATISTA P A, RONALDO C, *et al*. A study of the behavior of several methods for balancing machine learning training data[J]. SIGKDD Explorations, 2004, 6(1) : 20-29.

[15] TOMEK I. Two modifications of CNN[J]. IEEE Trans on Systems Man and Communications, 1976, 6: 769-772.

[16] HART P E. The condensed nearest neighbor rule[J]. IEEE Trans on Information Theory, 1968, 14(3) : 515-516.

[17] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C] //Proc of the 8th Conference on AI in Medicine. Europe: Artificial Intelligence Medicine, 2001: 63-66.

[18] WILSON D L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Trans on Systems, Man and Communications, 1972, 2(3) : 408-421.

(下转第 1308 页)