用于不均衡数据集的挖掘方法*)

赵凤英　王崇骏　陈世福
(南京大学计算机软件新技术国家重点实验室　南京 210093)
(南京大学计算机科学与技术系　南京 210093)

摘　要　传统的分类算法大多是基于数据集中各类的样本数是基本均衡的假设的，而实际应用场合中面临的往往是不均衡数据。针对不均衡数据集，利用传统的分类方法往往不能获得良好的性能，因而研究用于处理不均衡数据集的分类方法就显得相当重要，本文对相关的研究做了综述。

关键词　不均衡数据集，过取样，欠取样，代价敏感学习

Data Mining on Imbalanced Data Sets

ZHAO Feng-Ying　WANG Chong-Jun　CHEN Shi-Fu
(National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)
(Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

Abstract　The majority of machine learning algorithms previously designed usually assume that their training sets are well-balanced, but data in real-world is usually imbalanced. The tradition machine learning algorithms on balanced data sets have bad performance when they learn from imbalanced data sets. Thus, machine learning on imbalanced data sets becomes an urgent problem. In this paper, a simple review of the related work is informed.

Keywords　Imbalance data set, Over-sampling, Under-sampling Cost-sensitive learning

1　引言

分类是数据挖掘的重要目标之一，传统的分类方法大多基于如下假设[1]：(1)分类以高的正确率为目标；(2)数据集中各类的数目是均衡的；(3)各类的错分代价也是一样的。基于这样的假设研究出了很多分类算法，比如 ID3 算法、贝叶斯算法、神经网络等。

在很多实际应用场合中，如入侵检测、信用卡诈骗等，存在很多类别不均衡的情况。类别不均衡是指：在一个数据集中，一类样本数目特别多，而另一类的样本数目特别少，两者样本数目相差很大。比如在网络入侵检测中正常的网络访问要比入侵访问多得多；在保险业[2]，不论什么时候只会有一小部分保险客户要求索赔。医疗诊断，信用卡欺诈[3]等也存在这样的情况。传统的机器学习方法常常在这些不均衡数据集上对大类的分类性能较好，而对小类的分类性能却很糟糕。

小类别包含的信息不足以正确分类，是因为小类别的信息量无法与大类别相抗衡，其信息容易淹没在大类别中，导致小类别被大量误分。由于不均衡数据集中小类别的影响，均衡数据集的分类性能远远超过不均衡数据集的分类性能。几十年来人们提出了很多的学习类别不均衡问题的方法，本文对此做一个概要的综述，本文的结构如下：第 2 节介绍了若干学习类别不均衡问题的方法；第 3 节介绍了类别不均衡问题学习算法的评价标准；最后做了一个总结，并提出以后解决类别不均衡问题需要继续做的研究。

2　若干用于不均衡数据集的挖掘方法

用于学习类别不均衡数据集的方法主要有过取样方法、欠取样方法、代价敏感学习方法以及集成方法等。

2.1 过取样方法

过取样就是增加小类样本的数目。此方法的主要问题空间就是如何从现有的数据集中取样以及取多少。过取样方法通常复制一些小类样本，使小类的数目和大类的数目基本均衡。文[4]描述了用过取样方法如何将一个不均衡数据集均衡化。其步骤如下：

STEP1：初始的不均衡数据集；

STEP2：生成一些小类样本；

STEP3：用 Tomek links 作标记；

STEP4：去掉边界样本和噪音样本。

这里提及 Tomek link，Tomek link[5]，定义如下：

如果存在两个属于不同类别的样本 x 和 y，x 和 y 之间的距离用 $d(x,y)$ 表示，如果不存在另外一个样本 z 使得 $d(x,z)<d(x,y)$ 或 $d(y,z)<d(y,x)$，那么就称 (x,y) 为 Tomek link。

如果两个样本形成了一个 Tomek link，那么其中一个样本可能是噪音或者两个样本在边界上。虽然过取样方法通过复制小类样本可以均衡类别分布，但它可能会带来另外一个问题就是：大类样本可能会掺杂在小类样本也可能进入大类样本空间中，因此需要用 Tomek link 找出噪音和边界样本并将其去掉。

*)本文得到国家自然科学基金(No. 60503021)、江苏省自然科学基金(No. BK2005075)和江苏省高技术研究计划(No. BG2006027)的资助。**赵凤英**　硕士生，研究方向是机器学习与数据挖掘；**王崇骏**　博士，副教授，研究方向为机器学习与分布式人工智能；**陈世福**　教授，博士生导师，研究方向是人工智能与知识工程。

由于过取样方法复制了小类的样本，这容易带来过拟合问题。文[6,7]针对过拟合问题对过取样方法做了些改进，文[6]提出了一种新的方法 SMOTE，它不是简单地复制小类样本而是人工生成一些小类样本，从而避免了过拟合问题。

2.2 欠取样方法

欠取样就是减少大类样本的数目，此方法主要的问题空间就是如何产生样本以及产生多少样本。欠取样方法通常是去掉大类中的噪音、边界以及冗余数据样本，通常只能去掉一小部分。在大类和小类的比大于 100 时就不能有效地处理了，这时必须用集成子分类器的方法。欠取样的缺点是：它有可能去掉一些有用的样本，不能充分利用所给样本信息。文[8]用另外一种欠取样方法尽可能地不去掉有用的样本。大类样本被分为“安全样本”、“边界样本”和“噪音样本”。通常用 Tomek link 方法将边界样本和噪音样本从大类中去掉，只留下安全样本和小类样本作为分类器的训练集。

2.3 代价敏感学习方法

传统的分类器在学习不均衡数据集时，通常对大类分类的错误率较低，而对小类分类的错误率却很高。这种情况下如果把小类误分为大类的代价比把大类误分为小类的代价高很多时，尽可能地减少小类的错误率来降低误分类的总体代价就显得非常重要。当代价信息较容易得知时我们就可以用代价敏感学习方法来解决这个问题。

代价敏感学习方法是将各类不同的错分代价用到分类决策中去，目的是尽可能地降低误分类的总体代价而不是尽可能地降低误分类的错误率。C 类数据集的代价信息用代价矩阵 $Cost_{c\times c}$ 来表示。用 $Cost_{ij}$ 表示将第 i 个样本错分到 j 类的代价。正确分类的代价为零，将样本 x 错分为第 i 类的期望风险是 $R(i|x)=\sum_{j=1}^{c}P(j|x)Cost_{ij}$，在不知道样本 x 的真正类别时，将样本分为 i 类的风险用样本 x 的后验概率和错分代价来求得。对每个样本应将其分到期望风险最小的类别。

文[9]中提出了一个简单通用的方法可以使学习算法具有代价敏感性。基本思想是：改变训练集找那个的类别分布使其更倾向于代价高的类别。假如错分正类的代价是错分反类的代价的五倍，那么就把正类的样本数增加到原来的 5 倍。这样分类器就会倾向于正类分类的错误率。文[10]中提出了一个关于如何改变正类和反类占总样本数的比例来构造一个较优的代价敏感分类器的算法。

在现实中代价敏感学习方法通常难于确定代价的多少。比如，在医疗诊断中，即使我们知道小类错分的代价高于大类错分的代价，可是把癌细胞错分为健康细胞的代价应该设为多少？不过文[11]指出代价敏感学习和重取样方法取得了同样的效果。

2.4 类别均衡法

对于一个已经采集好的类别分布不均衡的训练集，为了尽量减少由于类别分布不均衡给分类性能带来的影响，文[12]提出了一种类别均衡法对类别分布不均衡的训练集进行处理。类别均衡法的本质就是使训练尽可能在数量级相当的类别上进行，避免对小类别的不公平对待。

类别均衡法的训练过程：

STEP1：先对训练集进行预处理，把所有的小类别合并成一个或几个新的较大的类别，这些新类别具有和训练集中原有的大类别相同或相近的数量级，由此形成一个新的类别分布比较均衡的训练集；

STEP2：在这个重新组合过的新训练集上进行训练，得到一个一级分类器；

STEP3：把原有的几个小类别组成一个小的训练集，进行分类训练，得到一个二级分类器。

使用类别均衡法会使训练时间增加，但不会增加太多，因为增加的时间是小类别样本训练时间，小类的样本数目较少，所以这种方法用较少的时间增加换来分类器性能的显著提高。

2.5 集成方法

集成方法通过联合许多学习器来以极小的代价获得很高的性能。欠取样方法通常不能充分利用数据集中大类样本。为了弥补它的不足，文[13]提出了两种集成方法（分别是 EasyEnsemble 和 BalanceCascade）来解决类别不均衡问题。EasyEnsemble 算法针对两类问题，假设大类为 N，小类为 P，从 N 中随机取得一个子集 N'，让 $|N'|=|P|$，同样用欠取样方法独立取多次，得到多个大类的子集，这些子集分别和小类样本结合起来作为训练集用来学习得到相应的分类器，再将这些分类器集成起来。BalanceCascade 算法的基本思想是：如果大类样本可以被原来得到的分类器正确分类，就视其为冗余的将其去掉，之后的取样是在剩下的样本空间中进行。重复此过程就可使训练样本达到基本均衡。文[13]实验证明这两种方法都能有效提高分类器的性能。

2.6 其他方法

除上面介绍的几种方法外还有很多别的解决方法。支持向量机是一个很重要的机器学习方法。它有很强的理论基础，也成功地应用到了很多领域，但在学习不均衡数据集时小类的错误率却很高。为了提高支持向量机的性能，文[14]做了一些研究，将 Class-Boundary-Alignment 方法应用到支持向量机中来使其更好地解决类别不均衡问题，得到两个新的算法 ACT 和 KBA。ACT 算法适用于可以用向量空间表示的数据集，KBA 适于不能用向量空间表示的数据集。文[14]将 Class-Boundary-Alignment 方法和 BM（Boundary Movement），BP（Biased Penalties）方法做了比较，发现：Class-Boundary-Alignment 表现出较好的性能。文[15]通过选择能有效分类小类的属性来改进 C4.5 算法并将其应用入侵数据集上取得了较好的效果。

另外无线电神经网络[16]、随机森林[17]、Mixture of Experts Agent[18]等等也是学习不均衡数据集的有效方法。在不同的应用领域，不同类型的类别不均衡问题往往需要采用不同的方法。

3 评价标准

混合矩阵是评价机器学习算法性能的典型方法[6]，如表 1 所示。

表 1 混合矩阵

	预测为反类的样本数	预测为正类的样本数
实际的反类样本数	TN	FP
实际的正类样本数	FN	TP

TN 是正确分为反类的样本数，FN 是把正类错分为反类的样本数，FP 是把反类错分为正类的样本数，TP 是正确分为正类的样本数。

所以：

正确率$=\frac{TN+TP}{TN+FN+FP+TP}$

错误率$=1-$正确率。

传统的分类器通常用错误率或正确率作为评价标准，旨在追求高的正确率。这种评价标准在不均衡学习问题上是显然不行的。比如，一个两类的不均衡数据集，大类的样本数占99%。用一个差的分类器把所有的样本都标记为大类，它的正确率就可以达到99%，在实际问题中通常很难有更好的分类器取得比这更高的正确率。如果以正确率为评价标准，这个分类器就算是做得很好了，但实际上并不是。在信用卡欺诈案件中，真正的欺诈案件是非常少的，按照这个做法，把所有的信用卡交易都归为正常的，这显然是不行的。在这些情况下通常用ROC(Receiver Operating Characteristic)图[19]作为分类器分类性能的评价标准。

针对两类问题的ROC图是一个真正率-伪正率的二维图。真正率$=\frac{TP}{FN+TP}$，伪正率$=\frac{FP}{TN+FP}$。X轴用伪正率表示，Y轴用真正率表示，画一张二维图，如图1所示。

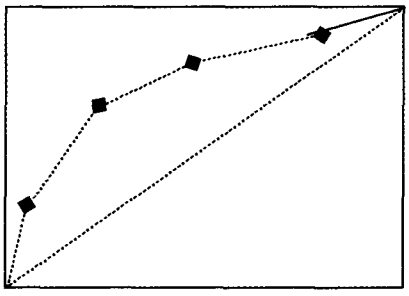

图1 ROC曲线

对角线代表随机分类器的分类性能。ROC曲线由联结一些离散的点而得，这些点是分类得到的结果。ROC曲线下面的面积称为AUC(Area Under the ROC Curve)。AUC可以用来比较分类器性能的优劣。如果一个分类器的AUC比另一个AUC大，则说明前者的分类性能比后者好。但对于某些有特殊的代价和类别分布的数据集，分类器能得到较大的AUC，但分类效果却不怎么令人满意。因此，有时候还需要用ROC凸壳方法(ROC convex hull)来衡量分类器的性能。

总结 不均衡数据集学习问题是机器学习中的一个很重要的问题，针对这一问题提出了很多的解决方法，一个普遍的方法是对不均衡的数据集用各种方法使其变得均衡。在很多实际问题中，通常存在两种情况：

(1)或者由太多的数据以至于算法处理不了，此时，必须从其中取出一部分样本来；

(2)或者根本就没有什么数据必须用一些方法产生数据。

目前已有的很多学习方法各有各自的优缺点，总体上来说，在处理不均衡问题上取得了不错的性能，但仍然存在一些问题需要解决，比如学习时遇到的噪音问题等；另外，大多数算法都是针对解决两类的类别不均衡学习问题的，但是在实际问题中存在着很多多类的类别不均衡问题，因此寻求解决多类问题的算法以及采用什么评价标准评价这样的算法也是需要继续研究的问题。

参考文献

1 Probost F. Machine learning from imbalanced data sets 101. In: the AAAI'2000 Workshop on Imbalanced Data Sets. 2000

2 Pednault E P D, Rosen B K, Apte C. Handling Imbalanced Data Sets in Insurance Risk Modeling: [Technical Report RC-21731]. IBM Research Re port, March 2000

3 Stolfo S J, Fan D W S, Lee W, et al. Prodromids, Chan P K. Credit Card Fraud Detection Using Meta-Learning: Issuesand Initial Results. In: AAAI-97 Workshop on AI Methods in Fraud and Risk Management, 1997

4 Batista G E A P A, Bazzan A L C, Monard M C. Balancing Training Data For Automated Annotation of Keywords: a Case Study. In: Proc. of the Second Brazilian Workshop on Bioinformatics. SBC, 2003

5 Tomek I. Two Modi_cations of CNN. IEEE Transactions on Systems Man and Communications, 1976, SMC-6: 769～772

6 Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P. SMOTE: Synthetic Minority Over-sampling Technique. Jounal of Artificial Intelligence Research, 2002, 16: 321～357

7 Batista G E A A, Carvalho A, Monard M C. Applying One-sided Selection to Unbalanced Datasets. In: O. Cairo, L. E. Sucar, and F. J. Cantu, eds. Proceedings of the Mexican International Conference on Artificial Intelligence- MICAI 2000, Springer-Verlag, Best Paper Award Winner, April 2000. 315～325

8 Kubat M, Matwin S. Addressing the Course of Imbalanced Training Sets: One-Sided Selection. In: Proceedings of 14th International Conference in Machine Learning, San Francisco, CA, 1997. 179～186

9 Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Wadsworth & Books, Pacific Grove, CA, 1984

10 Elkan C. The Foundations of Cost-Sensitive Learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, 2001. 973～978

11 Maloof M. Learning when data sets are imbalanced and when costs are unequal and unknown. ICML, 2003

12 张启蕊，张凌，董守斌，谭景华. 训练集类别分布对文本分类的影响. 清华大学学报(自然科学报)，2005，45(9)：1802～1805

13 Liu X-Y, Wu J, Zhou Z-H. Exploratory Under-Sampling for Class-Imbalance Learning. In: Proceedings of the 6th IEEE International Conference on Data Mining (ICDM'06), Hong Kong, China, 2006

14 Wu Gang. Class-Boundary Alignment for Imbalanced Dataset Learning. In: ICML-KDD'2003 Workshop: Learning from Imbalanced Data Sets, 2003

15 Zhou Quan, Gu Lin-Gang, Wang Chong-jun, et al. Using An Improved C4. 5 for Imbalanced Datatset of Intrusion. Privacy, Security, Trust, Ontario, Canada, 2003

16 Radivojac P, Korad U, Sivalingam K M, Obradovic Z. Learning from Class-Imbalanced Data in Wireless Sensor Networks. In: IEEE Semiannual Vehicular Technology Conference (VTC) Fall, (Orlando, FL), Oct. 2003

17 Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data: [Technical Report 666]. Statistics Department, University of California at Berkeley, 2004

18 Kotsiantis S B, Pintelas P E. Mixture of Expert Agents for Handling Imbalanced Data Sets - Kotsiantis, Pintelas, 2003

19 Fawcett T. ROC graphs: Notes and practical considerations for data mining researchers: [Technical Report HPL-2003-4]. HP Lab, Palo Alto, 2003