第11章：代价敏感学习

蒋良孝 |  中国地质大学（武汉）

机器学习与数据挖掘团队

ljiang@cug.edu.cn

http://grzy.cug.edu.cn/jlx/

本章内容

一、代价敏感学习的背景

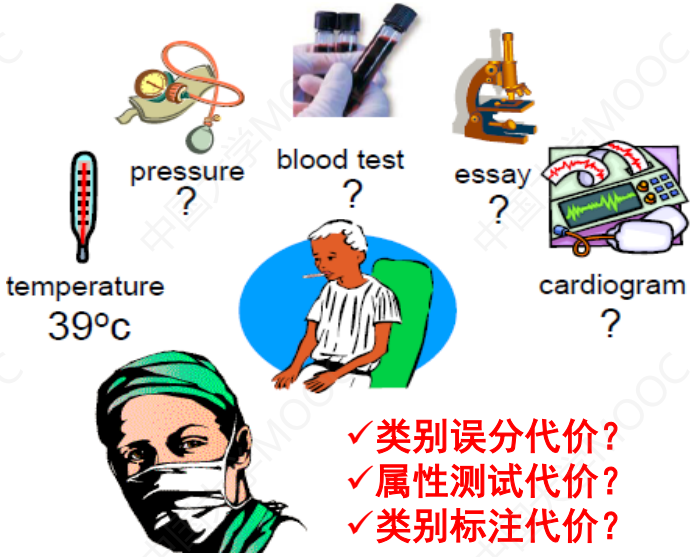
二、代价敏感学习的定义

三、代价敏感学习的评估

四、代价敏感学习的方法

一、代价敏感学习的背景

- 常用性能评估指标：准确率或错误率



- 传统的代价不敏感分类假定：
  1）不同类的误分类代价相同
  2）用于学习的训练数据足够完备
- 二者在许多实际的应用问题中很难得到满足。原因在于：
  1）不同类的误分类代价经常不同
  2）训练数据也经常因为获取困难、耗时、或者昂贵而不足
- 作为解决这些实际应用问题的关键技术，代价敏感学习被提出，并受到了机器学习与数据挖掘领域广大学者的高度重视

一、代价敏感学习的背景

International Journal of Information Technology & Decision Making
Vol. 5, No. 4 (2006) 597–604

World Scientific
www.worldscientific.com

10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH

QIANG YANG XINDONG WU

- Developing a unifying theory of data mining
- Scaling up for high dimensional data and high speed data streams
- Mining sequence data and time series data
- Mining complex knowledge from complex data
- Data mining in a network setting
- Distributed data mining and mining multi-agent data
- Data mining for biological and environmental problems
- Data Mining process-related problems
- Security, privacy and data integrity
- Dealing with non-static, unbalanced and cost-sensitive data

二、代价敏感学习的定义

- 代价敏感学习的目标是最小化分类器的总代价，而代价不敏感学习的目标是最小化分类器的错误率

- 分类器的代价主要体现在两个方面：
  1）误分类代价
  2）数据获取代价（属性测试代价和类别标注代价）

- 本课程主要关注误分类代价。误分类代价发生在分类器的预测阶段，主要包括：
  1）类依赖的误分类代价
  2）样本依赖的误分类代价

- 学习类依赖的误分类代价敏感的分类器要简单实用得多

二、代价敏感学习的定义

- 类依赖的误分类代价敏感的分类器采用最小化期望代价的原则来进行分类
- 分类器将一个待测样本 x 分成第 i 类的期望代价为：

$$R(i \mid x) = \sum_{j} P(j \mid x) C(i, j)$$

- 为了讲解方便，假定待分类的问题是一个二类分类问题，即只有正负两类，我们用1来表示正类，用0来表示负类，即二类分类问题的误分类代价矩阵如下：

真实情况	预测结果	
	正例	负例
正例	C (1,1)	C (0,1)
负例	C (1,0)	C (0,0)

三、代价敏感学习的评估

- 给定二类分类问题分类结果的混淆矩阵：

真实情况	预测结果	
	正例	负例
正例	TP（真正例）	FN（假负例）
负例	FP（假正例）	TN（真负例）

- 性能评估指标：误分类总代价

$$total\ misclassification\ costs = FN \times C(0,1) + FP \times C(1,0)$$

四、代价敏感学习的方法

Cost-Sensitive Learning

- Direct Cost-Sensitive Learning
  - Cost-senstive genetic algorithms, (ICET in short) (Turney, 1995)
  - Cost-senstive decision trees (Drummond and Holte, 2000; Ling et al., 2004)
  - Cost-senstive neural networks (Kukar and Kononenko,1998; Zhou and Liu, 2006; Chen et al., 2010)
  - Cost-senstive support vector machines (Morik et al., 1999; Fumera and Roli, 2002; Xu et al., 2006)
  - Cost-senstive Bayesian networks (Ibáñez et al., 2014)
  - Cost-sensitive Boosting (Fan et al., 1999; Shirazi and Vasconcelos, 2011)
- Cost-Sensitive Meta-Learning
  - Thresholding
    - ✓ MetaCost (Domingos, 1999)
    - ✓ CostSensitiveClassifier (CSC in short) (Witten and Frank, 2005)
    - ✓ Cost-sensitive naïve Bayes (Chai et al., 2004)
    - ✓ Empirical Thresholding (ET in short) (Sheng and Ling, 2006)
  - Rebalancing
    - ✓ Sampling (Jiang et al., 2013; Qiu and Jiang, 2015)
    - ✓ Weighting (Ting, 1998; Zadrozny et al., 2003; Jiang et al., 2014)
    - ✓ Cloning (Jiang et al., 2015)

四、代价敏感学习的方法

Direct Cost-Sensitive Learning

- Cost-senstive genetic algorithms, (ICET in short) (Turney, 1995)
- Cost-senstive decision trees (Drummond and Holte, 2000; Ling et al, 2004)
- Cost-senstive neural networks (Kukar and Kononenko,1998; Zhou and Liu, 2006; Chen et al, 2010)
- Cost-senstive support vector machines (Morik et al, 1999; Fumera and Roli, 2002; Xu et al, 2006)
- Cost-senstive Bayesian networks (Ibáñez et al, 2014)
- Cost-sensitive Boosting (Fan et al., 1999; Shirazi and Vasconcelos, 2011)

- 直接代价敏感学习方法直接将分类器的误分类总代价作为学习和优化的目标嵌入分类器的学习过程中，以使得分类器本身具有代价敏感性。

四、代价敏感学习的方法

Thresholding

- ✓ MetaCost (Domingos, 1999)
- ✓ CostSensitiveClassifier (CSC in short) (Witten and Frank, 2005)
- ✓ Cost-sensitive naïve Bayes (Chai et al., 2004)
- ✓ Empirical Thresholding (ET in short) (Sheng and Ling, 2006)

- 阈值调整的元学习方法通过调整分类器的判别阈值将代价不敏感的分类器转化成代价敏感的分类器。

- 代价敏感的分类器将一个待测样本 x 分成第 i 类，需要最小化期望代价：

$$R(i \mid x) = \sum_j P(j \mid x) C(i, j)$$

四、代价敏感学习的方法

- 根据这个公式，分类器将 x 分成正类，当且仅当：

$$P(0|x)C(1,0) + P(1|x)C(1,1) \leq P(0|x)C(0,0) + P(1|x)C(0,1)$$

$$P(0|x)(C(1,0) - C(0,0)) \leq P(1|x)(C(0,1) - C(1,1))$$

- 假定：$C(0,0) = C(1,1) = 0$
- 那么，上述公式可等价为：$P(0|x)C(1,0) \leq P(1|x)C(0,1)$
- 因为，$P(0|x) = 1 - P(1|x)$
- 所以，$P(1|x) \geq \dfrac{C(1,0)}{C(1,0)+C(0,1)}$

~~(0.5)~~

四、代价敏感学习的方法

Rebalancing

- ✓ Sampling (Jiang et al., 2013; Qiu and Jiang, 2015)
- ✓ Weighting (Ting, 1998; Zadrozny et al., 2003; Jiang et al., 2014)
- ✓ Cloning (Jiang et al., 2015)

- 再平衡的元学习方法通过再平衡训练样本集中不同类的误分总代价将代价不敏感的分类器转化成代价敏感的分类器。

- 正负两类样本再平衡的缩放比例：

$$p(1)\,C(0,1) : p(0)\,C(1,0)$$