计算机研究与发展 ISSN 1000-1239/CN 11-1777/TP
Journal of Computer Research and Development 47(9): 1539-1547, 2010

VOTCL及其在交叉销售问题上的应用研究

周广通 尹义龙 郭心建 董彩玲
(山东大学计算机科学与技术学院 济南 250101)
(zhouguangtong@gmail.com)

VOTCL and the Study of Its Application on Cross-Selling Problems

Zhou Guangtong, Yin Yilong, Guo Xinjian, and Dong Cailing
(*School of Computer Science and Technology, Shandong University, Jinan* 250101)

Abstract Cross-selling is regarded as one of the most promising strategies to make profits. The authors first describe a typical cross-selling model, followed by analysis showing that class-imbalance and cost-sensitivity usually co-exist in the data sets collected from this domain. In fact, the central issue in real-world cross-selling applications focuses on the identification of potential cross-selling customers. However, the performance of customer prediction suffers from the problem that class-imbalance and cost-sensitivity are arising simultaneously. To address this problem, an effective method called VOTCL is proposed. In the first stage, VOTCL generates a number of balanced training data sets by combining under-sampling and over-sampling techniques; then a base learner is trained on each of the data set in the second stage; finally, VOTCL obtains the final decision-making model by using an optimal threshold based voting scheme. The effectiveness of VOTCL is validated on the cross-selling data set provided by PAKDD 2007 competition where an AUC value of 0.6037 is achieved by using the proposed method. The ensemble model also outperforms a single base learner, which to some extent shows the efficacy of the proposed optimal threshold based voting scheme.

Key words cross-selling; class-imbalance; cost-sensitive; optimal threshold based voting; support vector machine

摘 要 交叉销售已成为企业盈利的重要手段,如何解决其数据中普遍同时存在的类别不平衡和代价敏感问题是准确预测交叉销售客户的关键,也是难点之一.针对上述问题,提出了一种基于最优阈值的投票方法:VOTCL.该方法首先结合过抽样和欠抽样技术获取多个类别平衡的训练数据集,然后在每个平衡数据集上分别训练得到多个底层学习器,最后利用所提出的基于最优阈值的投票集成方法集成底层学习器得到决策模型.在PAKDD 2007数据挖掘竞赛的交叉销售数据集上,VOTCL预测的AUC值为0.6037.该集成模型在性能上优于单个学习器,这也在一定程度上表明了所提出的基于最优阈值的投票集成方法的有效性.

关键词 交叉销售;类别不平衡;代价敏感;最优阈值投票;支持向量机

中图法分类号 TP181

收稿日期:2009-05-05;修回日期:2010-01-05

通信作者:尹义龙(ylyin@sdu.edu.cn)

0 引 言

类别平衡以及误分类代价相等是传统机器学习算法的必要前提[1], 但实际应用中该条件往往很难满足. 如在癌症检测问题中, 癌症患者在人群中所占比例只是很少一部分(类别不平衡性), 但误分类癌症患者所产生的代价要比误分类非癌症患者的代价高得多, 因为患者可能因此贻误治疗, 危及生命(误分类代价不同). 很多其他应用领域中也存在该类问题, 如电话诈骗检测[2]、雷达图像石油检测[3]等. 传统的机器学习算法力求全局的分类准确性, 往往将正类①中的数据分错, 造成高代价的误分结果. 因此, 类别不平衡和代价敏感问题已成为近年来机器学习领域的研究热点之一[4-5].

现有文献中对类别不平衡问题的研究工作主要集中在两个方面: 数据重抽样(resampling)和学习算法改进. 数据重抽样[5-7]是对数据集的操作, 包括过抽样(over-sampling)和欠抽样(under-sampling), 二者各有优劣, 过抽样增多了训练时可用的正类样本, 但会不可避免地引入一些噪声数据, 且学习的时间空间耗费也会随之增多; 欠抽样降低了学习耗费, 但同时又可能丢弃对分类潜在有用的样本. 学习算法改进则是通过改变算法内外部构造来处理类别不平衡问题的, 外部改进如阈值调整[5,8](主要针对学习算法的输出进行调整, 适当放宽对高代价正类样本的阈值限制)等, 内部改进针对具体算法作出改变, 如文献[9-10]涉及了对支持向量机算法的改进, 文献[11]是对决策树算法的改进等. 处理代价敏感问题时, 如误分类代价已知, 则学习问题相对简单, 基于最小代价的决策即可解决[4-5]; 但实际应用中, 误分类代价往往都是未知或无法确定的, 这就需要人为估计或给定代价(如文献[12]中 Heckman 算法), 然后通过重抽样或算法改进降低总体误分类代价, 如 MetaCost[13], Boosting[14], ETA[15]等.

交叉销售(cross-selling)[16-17]是指向已拥有本公司某业务的客户推销本公司的其他业务. 交叉销售问题可分为两类: 基于客户的交叉销售和基于业务的交叉销售. 现有文献[18-20]大都处理基于客户的交叉销售问题, 即利用机器学习和数据挖掘技术, 针对各个不同的客户, 发掘出不同客户群体对相应业务的不同需求, 从而指导交叉销售. 基于业务的交叉销售是面向特定业务的, 它通过对客户数据的学习, 预测出可能的参与该特定业务的客户, 从而指导交叉销售. 本文即针对基于业务的交叉销售问题展开研究.

与上述癌症检测和电话诈骗检测等应用类似, 交叉销售的数据中也同时存在类别不平衡问题和代价敏感问题(class-imbalance and cost-sensitive, 以下简称 CICS 问题), 此类问题一般表现为在现有的数个类别中, 有一个或者几个类别的样本个数相对稀少, 但是误分类这些样本所带来的代价又会比误分类其他样本高. 现有文献大都只针对单独的类别不平衡问题或代价敏感问题展开讨论, 而对 CICS 问题还没有一套比较成熟的解决方案. 文献[21-22]已经就类别不平衡性对学习算法的影响作了一些研究, 文献[21]中也提到同时考虑类别不平衡比率和误分类代价比率进行重抽样或阈值调整, 以解决 CICS 问题, 但实际应用中类别不平衡比率和误分类代价比率很难确定, 使得该方法的可操作性值得商榷.

为解决上述交叉销售中存在的 CICS 问题, 本文提出了一种有效的学习方法 VOTCL(voting based on an optimal threshold for class-imbalance and cost-sensitive learning). 首先, VOTCL 采用重抽样思想平衡两类数据, 即过抽样正类样本到一定数量, 随后欠抽样负类样本到相同数量并与过抽样后的正类数据集融合得到类别平衡的训练数据集, 此过程重复多次以得到数个不同的训练数据集; 其次, VOTCL 在每个训练数据集上分别训练得到底层学习器; 最后, 根据提出的基于最优阈值的投票集成方法, 并集成底层学习器得到最终决策模型. 在类别平衡数据集上的训练可以有效减弱类别不平衡性和代价敏感性对学习算法的影响; 基于最优阈值的投票集成方法可尽量避免误分正类样本, 从而有效降低了总体误分类代价. VOTCL 放宽了对类别不平衡比率和误分类代价比率的要求, 故该方法在可操作性和适应能力上有一定提高; 它本身也是一个封装方法(wrapper method), 可根据应用问题的不同而灵活选择底层学习器; 通过调整训练数据集的样本数量, 可有效控制模型训练时的时间空间需求, 这对处理实际应用中经常出现的海量数据有一定帮助. 此外, 实验结果也验证了 VOTCL 在处理交叉销售问题上的有效性.

① 如无特殊说明, 下面均指两类情况, 包括正类(少数类)和负类(多数类).

1 交叉销售

客户是企业生存和发展的根基, 而保护原有客户、吸引新客户是企业提高核心竞争力的关键所在. 企业向有价值的客户提供交叉服务不仅可以提升客户价值, 扩大自身收入及利润, 而且还能够提高客户的满意度、忠诚度, 增强企业竞争力. 因此, 通过利用机器学习和数据挖掘技术分析客户信息, 充分挖掘客户数据中潜在的知识或规则, 可有效地根据不同客户的偏好和特性提供相应的产品和服务, 进行交叉销售, 实现企业和客户的双赢.

1.1 问题描述

设某公司可提供多种业务 $A=\{A_1, A_2, \cdots, A_n\}$, 公司待分类客户集 $C=C_1 \cup C_2 \cup \cdots \cup C_n$, 并且相应于业务 A_i 的客户集为 C_i(客户集之间可能会有交叉和重叠);

如果公司准备向拥有业务 A_i 的某些客户推销本公司的另一种业务 A_j, 则需要对 C_i 中的客户作出测评, 以确定可能参与业务 A_j 的客户, 公司也可据此向这些客户宣传业务 A_j, 进行交叉销售.

1.2 问题分析

上述交叉销售问题的关键在于对客户测评, 这也可被视为一个机器学习的过程. 其中训练数据是公司的历史客户集 $C_h=C_{h1} \cup C_{h2} \cup \cdots \cup C_{hm}$, 它描述过去某段时间内公司的客户信息. 如果公司准备向拥有业务 A_i 的客户推销业务 A_j, 则 C_{hi} 中应该有一部分客户既拥有业务 A_i 也拥有业务 A_j. 据此可把 C_{hi} 分为两类, $C_{hi}=C_{hi}^{+} \cup C_{hi}^{-}$, 既拥有业务 A_i 也拥有业务 A_j 的客户组成正类(C_{hi}^{+}), 此类客户本身就拥有多重业务, 其接受新业务的可能性相对较大, 因此在交叉销售中应受到更多关注; 其他客户只是参与了业务 A_i, 参与新业务的可能性较小, 这些样本组成负类(C_{hi}^{-}). 同样, 待分类客户集(也即预测数据) C_i 中也应该存在正负两类: $C_i=C_i^{+} \cup C_i^{-}$, 学习的任务就是在待分类客户集中尽可能精确而且全面地预测出交叉销售客户 C_i^{+}, 进而指导交叉销售.

交叉销售中同时存在类别不平衡问题和代价敏感问题, 是一个典型的 CICS 问题.

1. 类别不平衡问题: 实际应用中, 同时参与业务 A_i 和 A_j 的客户是相对稀少的, 即 $|C_{hi}^{+}| \ll |C_{hi}^{-}|$ 且 $|C_i^{+}| \ll |C_i^{-}|$, 历史客户集和待分类客户集中存在类别不平衡性.

2. 代价敏感问题: 误分一个交叉销售客户(正类客户)带来的代价远大于误分一个负类客户带来的代价, 因为误分负类的客户的代价只是一些宣传成本, 而误分类正类客户的代价将有可能是丢失销售机会甚至巨额的业务利润, 并且该误分类代价很难事先确定.

因此, 在对交叉销售问题进行分析时应当充分考虑类别不平衡性和代价敏感性的影响, 并有效解决该问题; 测评时尽量少地误分类高代价正类客户, 以尽可能提高交叉销售的盈利效果.

2 基于最优阈值的投票方法 VOTCL

为有效解决交叉销售应用中出现的 CICS 问题, 本文提出了一种基于最优阈值的投票方法 VOTCL. VOTCL 首先采用重抽样思想平衡两类数据并减少训练样本数, 该过程重复多次得到多个训练数据集, 随后 VOTCL 基于这些平衡数据集训练出多个底层学习器, 最后利用基于最优阈值的投票方法集成底层学习器, 得到最终决策模型. VOTCL 的总体框架如图 1 所示. 详细的算法流程如算法 1 所示.

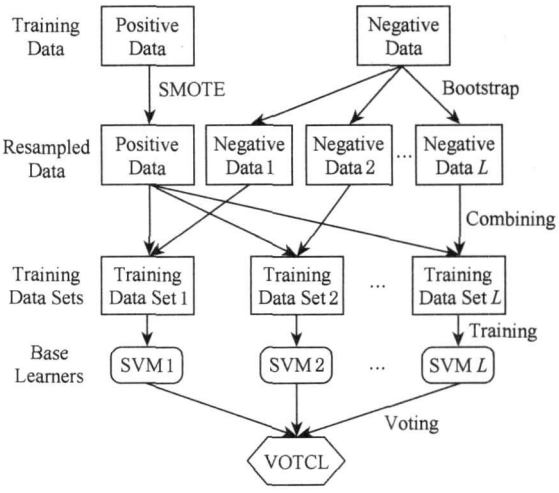

Fig. 1 Framework of VOTCL.

图 1 VOTCL 总体框架图

算法 **1.** VOTCL 算法流程.

输入: D: 训练数据, 如 C_{hi};

P: 待预测数据, 如 C_i;

L: 底层学习器个数.

进程:

① $D^{+} \leftarrow D$ 中的正类样本;

② $D^{-} \leftarrow D$ 中的负类样本;

③ $D_{\mathrm{SMOTE}}^{+} \leftarrow \mathrm{SMOTE}(D^{+})$;

④for $i=1$ to L do

⑤ $D_i^- \leftarrow \text{Bootstrap}(D^-)$;

⑥ $D^i \leftarrow D_i^- \cup D_{\text{SMOTE}}^+$;

⑦ $SVM_i \leftarrow$ 在 D^i 上训练 SVM 模型;

⑧end

输出:

① $v(x)=\sum_{i=1}^{L} SVM_i(x)$;

②根据 F 度量选取最优的阈值 K;

③if $v(x) \geqslant K$ then

④ $Label(x)=1$;

⑤ $R(x)=\frac{1}{v(x)}\sum_j prob_j^+(x)$, $j \in \{j \mid SVM_j(x)=1\}$;

⑥else

⑦ $Label(x)=0$;

⑧ $R(x)=\frac{1}{L-v(x)}\sum_k prob_j^+(x)$, $k \in \{k \mid SVM_k(x)=0\}$;

⑨end

⑩return $Label(x)$ 和 $R(x)$.

2.1 底层学习算法选择

本文选用支持向量机[23](support vector machine, SVM)作为底层学习器. 支持向量机训练完成后, 所得 SVM 模型可用于分类(给出类别标号)和回归(给出后验概率). 对于两类问题中某样本 x, 模型预测标号设为 $SVM(x)$(取值 0 或者 1, 0 为负类, 1 为正类), 后验概率设为 $prob^+(x)$(样本 x 被划分为正类的后验概率)和 $prob^-(x)$(样本 x 被划分为负类的后验概率).

2.2 数据重抽样

文献[21]通过大量实验得出, 当误分类代价差别不大并且类别不平衡率较小时, 使用原始数据学习可获得较好效果; 但在误分类代价差别较大或类别严重不平衡的情况下, 应首先平衡两类数据, 然后再进行学习. 数据重抽样是最直接的平衡数据的方法, 而 SMOTE[6]和 Bootstrap[7]则分别是目前最常用的过抽样和欠抽样方法. 文献[24]指出, SMOTE 能够起到一定平滑 SVM 决策边界的作用. 但由于交叉销售问题一般都存在严重的类别不平衡性, 仅仅使用 SMOTE 平衡两类样本会产生过多的正类样本, 数据量也会随之增大; 并且, 由于正类样本往往都比较接近决策边界, 这些新增样本又不可避免地具有增加分类噪声的危险. 因此, 我们试图通过结合过抽样和欠抽样技术来得到类别平衡的数据集.

VOTCL 首先使用 SMOTE 过抽样正类样本到一定数量 N_{SMOTE}($N_{\text{SMOTE}} < N$, N 为训练样本总数, 算法 1 中过程③), 然后应用 Bootstrap 欠抽样负类样本到相近的数量 $N_{\text{Bootstrap}} \approx N_{\text{SMOTE}}$(算法 1 中过程⑤)并在融合后(算法 1 中过程⑥)形成训练数据集, 这样两类样本也得以平衡并且训练数据集的样本数量($N_{\text{Bootstrap}}+N_{\text{SMOTE}}$)较训练样本总数 N 也有效减少; 重复 L 次 Bootstrap 并分别与 SMOTE 后的数据集融合即可得到 L 个类别平衡的训练数据集(算法 1 中过程④~⑧). 该重抽样方法可以有效控制参与训练的样本数量, 并且底层学习器的训练也可并行进行, 使得训练过程的时间、空间消耗能够得以控制, 这对处理实际应用中经常出现的海量数据有一定帮助. 类似的抽样策略也出现在之前的文献中[24].

2.3 基于最优阈值的投票集成方法

令 $SVM_j(x)$ 表示第 j($j \in \{1, 2, \cdots, L\}$)个 SVM 对样本 x 的预测标号, $prob_j^+(x)$ 表示第 j 个 SVM 预测样本 x 为正类的后验概率, 则基于最优阈值的投票集成方法的分类函数和回归函数可分别定义如下.

分类函数 $Label(x)$: 首先统计将样本 x 预测为正类的底层学习器个数, 即投票数 $v(x)$(算法 1 中输出①). 如果有 K 个或多于 K 个底层学习器预测样本 x 为正类, 则 VOTCL 最终预测 x 为正类(算法 1 中输出④), 否则为负类(算法 1 中输出⑦). 算法可通过在 1~L 之间调整阈值 K, 以尽量在不误分正类的情况下提高负类的分类正确率; 此处性能评价标准选用 F 度量(F-measure)(算法 1 中输出②). 由于 F 度量中召回率的提高意味着正类误分比率的降低, 在处理交叉销售问题中, 我们有意设置 F 度量中的参数 $\beta<1$, 即增强召回率所占的权重, 从而利用这种集成方式有效降低算法的总体误分类代价, 这也是基于最优阈值投票思想的核心.

回归函数 $R(x)$: 如 VOTCL 预测该样本为正类, 则其回归值(也即本文中的得分)定义为所有预测 x 为正类的底层学习器的后验概率的均值(算法 1 中输出⑤); 如 VOTCL 预测该样本为负类, 则其回归值则由所有预测 x 为负类的底层学习器的后验概率平均得到(算法 1 中输出⑧), 此处涉及的后验概率均为正类后验概率.

3 实 验

3.1 交叉销售数据集

本实验基于PAKDD 2007 数据挖掘竞赛①提供的交叉销售数据集, 该数据是由某金融公司提供的真实应用中的数据. 训练数据 $C_{h\_credit\_card}$ 中的所有客户均开通了信用卡业务, 共有 40 700 个样本, 属性维数为 40. 其中正类样本集 $C^{+}_{h\_credit\_card}$ 中有 700 个样本, 相应客户在开通信用卡业务之后又参与了贷款业务; 其余 40 000 个客户只开通了信用卡业务, 他们组成了负类样本集 $C^{-}_{h\_credit\_card}$. 预测数据 C_{credit_card} 共 8 000 个客户样本, 属性维数同样为 40. 该交叉销售的任务就是为 C_{credit_card} 中的每个客户样本生成得分, 得分越高表示该客户参与贷款业务的可能性越大.

该数据集中存在的主要问题有:

1. 类别不平衡问题. 该交叉销售应用是一个两类问题, 并且在给定的训练数据中存在严重的类别不平衡, 正类样本数和负类样本数之比为 40 000∶700;

2. 代价敏感问题. 公司通过贷款业务获取的利润明显要比通过信用卡业务所能获取的利润高, 误分类贷款客户的代价远大于误分类信用卡客户的代价;

3. 得分问题. 公司希望为预测数据中的客户样本生成得分, 该得分可以反映客户参与贷款业务的倾向性大小. 公司期望生成的理想结果是: 为 $C^{-}_{credit\_card}$ 中的客户生成较低的得分, 而为 $C^{+}_{credi\_cardi}$ 中的客户生成较高的得分, 并且两者之间没有交叉重叠, 如果能达到这样的理想结果, 该金融公司便可根据得分由高到低地向客户宣传贷款业务. 因此, 得分的精确性和序列性需要保证.

该数据中存在的其他问题有:

1. 训练数据有40 700 个样本, 且每个样本都有 40 个不同的属性, 数据量较大;

2. 训练数据和预测数据中均存在大量缺失数据和描述型数据. 有效处理这些问题也是非常重要的.

3.2 实验设计及分析

3.2.1 预处理和数据重抽样

数据集中的缺失数据主要集中在 7 个属性上, 如表 1 所示. 此外, 数据中还存在多个描述型属性, 为实施 SVM 算法, 需要把描述型属性转化为数值型属性. 实验对训练数据和预测数据中的空缺值的采用类条件概率[25] 填充, 对描述型属性用全局平均[26] 方法转化为数值型属性. 最后再将所有属性值归一化到[0, 1]区间内.

Table 1 Statistical Results of Missing Values in the Training and Prediction Data

表 1 训练数据和预测数据中的缺失统计

Attribute	Training Data		Prediction Data	
	Amount	Proportion /%	Amount	Proportion /%
AMEX_CARD	3707	9.09	721	9.01
DINERS_CARD	3884	9.52	760	9.50
VISA_CARD	2455	6.02	477	5.96
MASTER_CARD	2937	7.20	570	7.13
RETAIL_CARDS	3693	9.05	724	9.05
DISP_INCOME_CODE	28742	70.45	5672	70.90
CUSTOMER_SEGMENT	1740	4.26	338	4.23

根据 2.2 节的分析, VOTCL 需要采用重抽样方法平衡两类数据, 并得到多个训练数据集. 本实验中的实现如下: 首先使用 SMOTE 过抽样训练正类样本到 N_{SMOTE} = 2 100 个, 然后 Bootstrap 训练负类样本到 2 100 个, 重复 Bootstrap 过程 20(L = 20)次并分别与SMOTE 后的数据集融合, 得到20 个不同的平衡训练数据集. 此处 N_{SMOTE} 和 L 是经过多次实验得到的适合该数据集的经验值, 其中我们也参考了前人的工作结果(正如 2.2 节所介绍的), 并考虑了算法在实验数据集上的时间开销.

3.2.2 底层学习器训练

训练底层学习器时, 由于存在代价敏感问题, 错误率已经不能作为学习器的性能评测标准[4-5, 12], AUC[27] 值已被广泛应用到类别不平衡学习和代价敏感学习的评测中, 并且实践也证明其良好的性能②. 实验中选用 AUC 值作为底层 SVM 算法的性能评测标准以选取参数.

本实验基于 LIBSVM[28] 实现 SVM 算法, 其中核函数选用径向基(RBF) 核函数. 为训练底层 SVM 学习器, 需调整径向基核函数中的参数 γ 和 SVM 的惩罚因子 C. 本实验中采取的策略如下: 我们首先选取一个较大的参数空间, 训练得到其相应的 AUC 值, 如图 2(a) 所示(该结果为训练数据集 1 上的

① PAKDD 2007 数据挖掘竞赛的主页地址为 http://lamda.nju.edu.cn/conf/pakdd07/dmc07/index.htm.

② 事实上, PAKDD 2007 数据挖掘竞赛也采用 AUC 值作为最终结果的评测标准.

结果); 根据图 2(a), 我们可选定 AUC 值较高的小参数区间, 细化参数并继续训练 SVM, 获取其 AUC 值, 如图 2(b) 所示; 最后根据图 2(b) 即可选取优化参数(图 2 例子中的优化参数取为 $C=5$, $\gamma=21$).

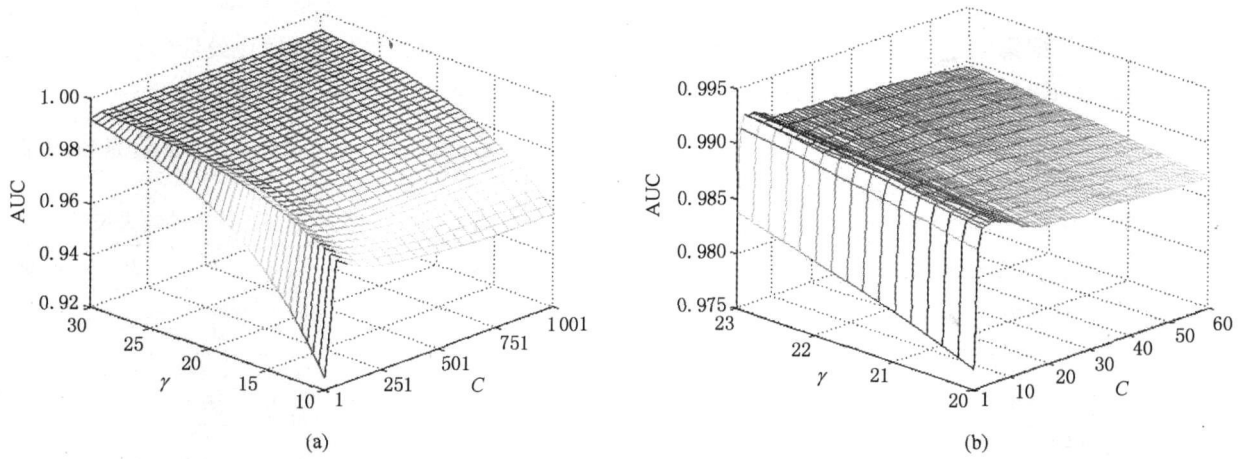

Fig. 2 Parameter setting for SVM on training data set 1. (a) AUC values of rough parameter settings and (b) AUC values of the refined parameter settings.

图 2 训练数据集 1 上 SVM 的参数选择. (a) 大参数空间上的 AUC 值; (b) 细化参数空间上的 AUC 值

3.2.3 集 成

根据基于最优阈值的投票集成方法, 实验首先计算阈值 K 在不同取值(从 1~20) 时正负两类的预测正确率, 结果如图 3(a) 所示; 之后计算相应的 F 度量值(参数 β 设为 0.1), 结果如图 3(b) 所示. 由图 3(b) 可知, 当 K 取值 15 时, F 度量值可取最大(0.999 79), 故本实验中选取最优阈值 $K=15$ 对底层学习器集成, 并构建最终决策模型.

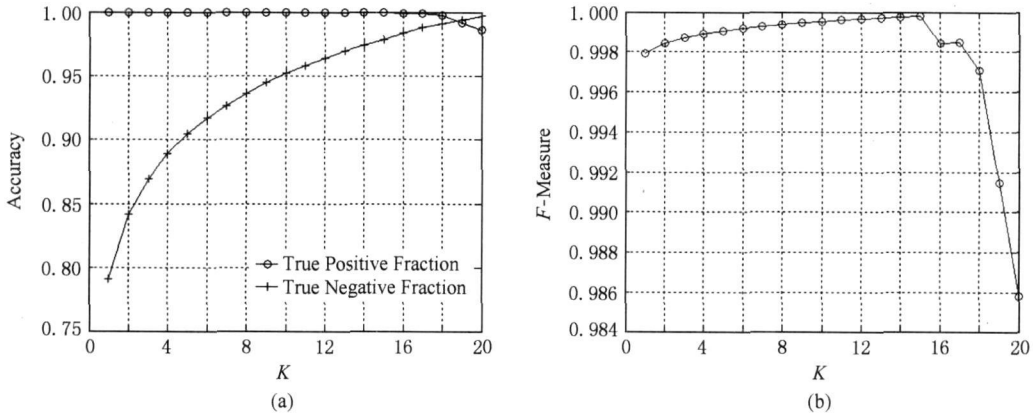

Fig. 3 Selection of the optimal threshold K. (a) True positive/negative fraction for different K values and (b) F-measure for different K values.

图 3 最优阈值 K 的选择. (a) 不同 K 值所对应的正负两类预测正确率; (b) 不同 K 值所对应的 F 度量值

3.2.4 预测和得分

图 4 是 VOTCL 对训练数据的得分分布情况. 对于每个样本得分越高说明其参与贷款业务的可能性越大, 因此我们以 0.5 时作为阈值, 得分大于 0.5 时被认定为正类, 否则为负类. 图 4 中黑色点为正类样本, 深灰色点为误分的负类样本, 浅灰色点为正确分类的负类样本. 其中, 所有的正类样本都正确预测, 而且只有很少比例的负类样本(2.15%) 被误分, 该得分保证了较高的精确性; 正类样本的得分普遍高于其余样本, 少数负类样本虽然被错误预测, 但其得分也明显低于正类样本, 得分的序列性也较好地得以保证.

VOTCL 在 PAKDD 2007 的交叉销售问题上所得最终 AUC 值为 0.6037(由 PAKDD 2007 竞赛组委会给出①, 该竞赛冠军的 AUC 值为 0.7001). 此外, 此外, 单独使用底层 SVM 学习器所预测的 AUC 值为

① 最终排名见 http://lamda.nju.edu.cn/conf/pakdd07/dmc07/results.htm, 本文方法的参赛编号为 P060.

0.591 1(在 20 个底层学习器上的平均结果), 方差为 0.000 09, 此结果也进一步说明了本文所提出的最优阈值投票集成的有效性.

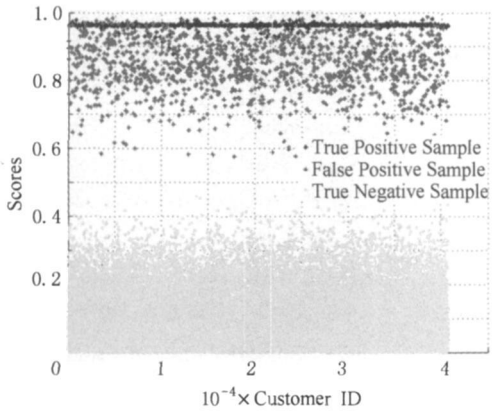

Fig. 4 Distribution of scores for the samples in training data.

图 4 训练数据上的得分分布情况

3.3 VOTCL 决策机理解析

由于 VOTCL 在 L 个不同的训练数据集上得到 L 个不同的SVM, 每个SVM的决策面都可以被视为高维空间上的一个超平面, 这些超平面是由多次 Bootstrap 欠抽样后的负类样本分别与 SMOTE 过抽样后的正类样本融合并训练得到的. 选取最优阈值 K 意味着把至少 K 个超平面组合起来, 这里每种超平面的组合都可以分割出一个子空间, 这些子空间的并集也组成了最终 VOTCL 的决策空间. 这也类似于分段线性分类器, 只是分段线性分类器的决策空间是静态固定的, VOTCL 超平面的组合是动态变化的, 大于等于 K 个的超平面组合所形成的子空间都可用来决策, 因此 VOTCL 也适用于较为复杂的决策情况.

图 5 中的例子可以更好地说明 VOTCL 的决策机理. 该例子中有 4 个决策线, 即 l_1, l_2, l_3, l_4, 决策线上箭头所指方向为相应的正类决策空间, 这 4 条决策线把整个样本空间划分为 A ~ J 共 10 个子空间. 如选取最优阈值 K= 3 来投票集成这 4 条决策线, 则由 3 条或 3 条以上的决策线共同决策的区域都可认定为正类的分布空间, 3 条决策线共同决策区域可见图 5(b) ~ (e), 4 条决策线共同决策区域可见图 5(f), 最终组合后的 VOTCL 决策空间可见图 5(g), 图中黑框字母所标示的区域代表正类样本的决策空间:

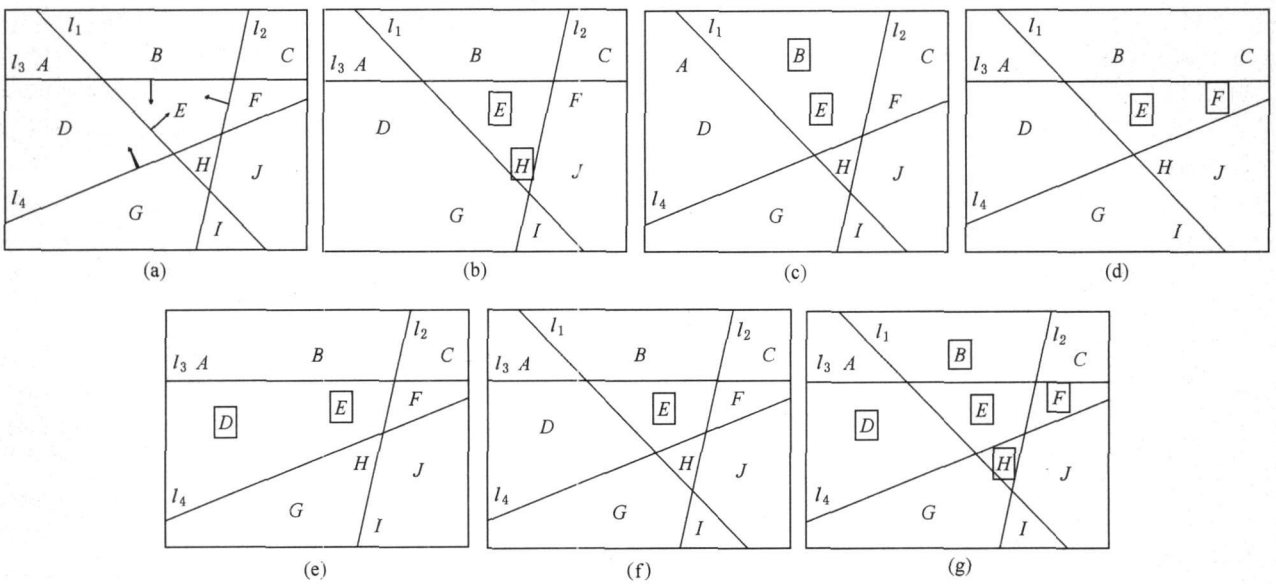

Fig. 5 Analysis of the decisionr making of VOTCL. (a) Sample space and decisionlines $l_1l_2l_3$ and l_4; (b) Joint decision space of l_1l_2 and l_3; (c) Joint decision space of l_1l_2 and l_4; (d) Joint decision space of l_1l_3 and l_4; (e) Joint decision space of l_2l_3 and l_4; (f) Joint decision space of $l_1l_2l_3$ and l_4; and (g) Decision space of VOTCL.

图 5 VOTCL 决策机理解析. (a) 样本空间及决策线 $l_1l_2l_3l_4$; (b) $l_1l_2l_3$ 的共同决策空间; (c) $l_1l_2l_4$ 的共同决策空间; (d) $l_1l_3l_4$ 的共同决策空间; (e) $l_2l_3l_4$ 的共同决策空间; (f) $l_1l_2l_3l_4$ 的共同决策空间; (g) VOTCL 的决策空间

4 结论与下一步工作

本文研究和分析了目前企业投资提高利润的一个核心策略——交叉销售, 探讨了其数据集中同时存在的类别不平衡性和代价敏感性, 并提出使用 VOTCL 来有效处理该问题. VOTCL 通过重抽样的办法获得较为平衡的训练数据集, 可有效减弱类别不平衡性和代价敏感性对底层学习器的影响. 基于最优阈值的投票集成方法在一定程度上避免了正类样本

的误分, 有效降低了总体误分类代价. 本文在 PAKDD 2007 数据挖掘竞赛提供的交叉销售数据上的实验结果也验证了 VOTCL 的有效性. 总体而言, VOTCL 放宽了对类别不平衡比率和误分类代价比率的要求, 使得其可操作性和适应能力有一定提高. 通过对 VOTCL 决策机理的进一步分析可知, 基于最优阈值的投票集成方法能够动态集成多种超平面组合的决策结果, 因此该方法可适用于较为复杂的决策情况. VOTCL 采用的重抽样方法可有效控制参与训练的样本数量, 且底层学习器的训练也可并行进行, 故本方法在处理海量数据时有一定优势. 此外, VOTCL 本身也是一个封装方法, 可根据应用问题灵活选择底层学习器.

VOTCL 在其他问题上的应用有待于进一步研究. 底层 SVM 参数的自动优化和选择也是今后工作的方向. 此外, VOTCL 中参数(L 和 N_{SMOTE})的选择在一定程度上会影响最终算法的性能, 后续工作也将就此展开研究.

参 考 文 献

[1] Duda R, Hart P E, Stork D. Pattern Classification [M]. New York: Wiley, 2000

[2] Fawcett T, Provost F. Adaptive fraud detection [J]. Data Mining and Knowledge Discovery, 1997, 1(3): 291-316

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

[4] Viaene S, Dedene G. Cost-sensitive learning and decision making revisited [J]. European Journal of Operational Research, 2005, 166(1): 212-220

[5] Elkan C. The foundations of cost-sensitive learning [C] //Proc of IJCAI'01. San Francisco: Morgan Kaufmann, 2001: 973-978

[6] Chawla N, Bowyer K, Hall L, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357

[7] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140

[8] Weiss G. Mining with rarity: A unifying framework [J]. SIGKDD Explorations, 2004, 6(1): 7-19

[9] Lessmann S. Solving imbalanced classification problems with support vector machines [C] //Proc of ICAI'04. Las Vegas, NV: CSREA Press, 2004: 214-220

[10] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines [C] //Proc of IJCAI'99. San Francisco: Morgan Kaufmann, 1999: 55-60

[11] Chen Chao, Liaw A, Breiman L. Using random forest to learn imbalanced data, 666 [R]. Berkeley: University of California at Berkeley, 2004

[12] Zadrozny B, Elkan C. Learning and making decisions when costs and probabilities are both unknown [C] //Proc of ACM SIGKDD'01. New York: ACM, 2001: 204-213

[13] Domingos P. MetaCost: A general method for making classifiers cost-sensitive [C] //Proc of ACM SIGKDD'99. New York: ACM, 1999: 155-164

[14] Fan Wei, Stolfo S, Zhang Junxin et al. AdaCost: Misclassification cost-sensitive boosting [C] //Proc of ICML'99. San Francisco: Morgan Kaufmann, 1999: 97-105

[15] Charles L, Victor S. A comparative study of cost-sensitive learning [J]. China Journal of Computers. 2007, 30(8): 1203-1212 (in Chinese)
(Charlesl, Victor S. 代价敏感分类器的比较研究[J]. 计算机学报, 2007, 30(8): 1203-1212)

[16] Kamakura W, Kossar B, Wedel M. Identifying innovators for the cross-selling of new products [J]. Management Science, 2004, 50(8): 1120-1133

[17] Berson A, Smith S, Thearling K. Building Data Mining Applications for CRM [M]. New York: McGraw-Hill, 1999

[18] Li Shibo, Sun Baohong, Wilcox R. Cross-selling sequentially ordered products: An application to consumer banking services [J]. Journal of Marketing Research, 2004, 42(2): 233-239

[19] Wong R, Fu A, Wang Ke. Data mining for inventory item selection with cross-selling considerations [J]. Data Mining and Knowledge Discovery, 2005, 11(1): 81-112

[20] Bhasker B, Park H, Park J, et al. Product recommendations for cross-selling in electronic business [G] //LNCS 4304: Proc of AUSAI'06. Berlin: Springer, 2006: 1042-1047

[21] Liu Xuying, Zhou Zhihua. The influence of class imbalance on cost-sensitive learning: An empirical study [C] //Proc of ICDM'06. Los Alamitos, CA: IEEE Computer Society, 2006: 970-974

[22] Weiss G, Provost F. The effect of class distribution on classifier learning, ML-TR 43 [R]. New York: Rutgers University, 2001

[23] Vapnik V. The Nature of Statistical Learning Theory [M]. Berlin: Springer, 1995

[24] Liu Yang, An Aijun, Huang Xiangji. Boosting prediction accuracy on imbalanced datasets with SVM ensembles [G] //LNCS 3918: Proc of PAKDD'06. Berlin: Springer, 2006: 107-118

[25] Scheffer J. Dealing with missing data [J]. Research Letters in the Information and Mathematical Sciences, 2002, 3: 153-160

[26] Li Cen, Biswas G. Unsupervised learning with mixed numeric and nominal data [J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(4): 673-690

[27] Bradley A. The use of the area under the ROC curve in the evaluation of machine learning algorithms [J]. Pattern Recognition, 1997, 30(7): 1145-1159

[28] Chang C, Lin C. LIBSVM: A library for support vector machines [OL]. [2007-04-10]. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Zhou Guangtong, born in 1986. Received his BS degree from Shandong University in 2007. Now he is a master candidate at the School of Computer Science and Technology, Shandong University. His main research interests include machine learning, data mining and their applications to biometrics and content-based image retrieval.

周广通, 1986 年生, 硕士研究生, 主要研究方向为机器学习、数据挖掘及其应用.

Yin Yilong, born in 1972. Received his PhD degree from Jilin University in 2000, and now he is a professor and PhD supervisor in the School of Computer Science and Technology, Shandong University. He is a committee member of Chinese Association of Artificial Intelligence (CAAI), executive committee member of the CAAI Machinelearning Society, member of the China Computer Federation (CCF). His main research interests include machine learning and applications, biometrics.

尹义龙, 1972 年生, 博士, 教授, 博士生导师, 中国人工智能学会理事、机器学习专委会常务委员、中国计算机学会会员, 主要研究方向为机器学习及应用、生物特征识别.

Guo Xinjian, born in 1980. Received his BS degree from Shandong University in 2007, and now he is a master candidate at the School of Computer Science and Technology, Shandong University. His main research interests include machine learning, machine version and their applications.

郭心建, 1980 年生, 硕士研究生, 主要研究方向为机器学习、机器视觉及其应用.

Dong Cailing, born in 1983. Received her BS degree from Shandong University in 2007, and now she is a master candidate at the School of Computer Science and Technology, Shandong University. Her main research interests include data mining and its applications, machine learning.

董彩玲, 1983 年生, 硕士研究生, 主要研究方向为数据挖掘及其应用、机器学习.

Research Background

Class-imbalance and cost-sensitivity usually coexist in real-world cross-selling data collections, and the performance of identifying potential cross-selling customers suffers from these problems. In fact, the problems of class-imbalance and cost-sensitive have already been noticed and studied in machine learning community, and the past years also witnessed a substantial progress in the development of fast and effectivelearning techniques. However, despite the various methods dealing with class-imbalance and cost-sensitivity separately, there are few successful approaches which address the problem with coexisting class-imbalance and cost-sensitivity. Most previous works suggest resampling the data sets or adjusting the thresholds based on class-imbalance ratio and cost-sensitive ratio, but it is unpractical to estimate the two ratios in real-world applications. To address these problems, we propose VOTCL for effectively predicting potential cross-selling customers. VOTCL first combines under-sampling and over-sampling techniques to obtain balanced training data sets, based on which a number of base learners are trained. We ensemble the base learners with an optimal threshold based voting scheme, and the final decision is made by the ensemble model. The effect of class-imbalance and cost-sensitivity is weakened by training based on balanced data sets. The proposed optimal threshold based voting also helps to improve prediction performance, as compared to a single base learner. VOTCL relaxes the requirements for class-imbalance ratio and cost-sensitive ratio. Useful insight is provided through the analysis of the decision-making principles of VOTCL.