基于 Boosting 的市场值函数算法及其评价

刘椿年，袁彩卿，黄佳进，欧创新
（北京工业大学 计算机学院 多媒体与智能软件北京市重点实验室，北京 100022）

摘 要：发现具有潜在市场价值的客户群是直销中的一个关键问题，尽管一些标准的数据挖掘算法可以用来解决此问题，但效果并不理想．为此，采用市场值函数算法，它以信息论为基础，通过构造一个线性市场值函数来对客户进行排序，从而发现最具有潜在市场价值的客户群．实验结果表明，它的评价值可达 80% 以上，并且具有很好的可解释性．同时，将 Boosting 算法应用到市场值函数算法中，用以提高市场值函数的预测效果；在 3 个不同的数据集上进行了实验，评价值均提高了一个百分点．

关键词：Boosting 算法；市场值函数；lift 值
中图分类号：TP 39 **文献标识码**：A **文章编号**：0254－0037(2004)03－0369－04

商业竞争中，如果商家能够发现具有潜在市场价值的客户群，只向该客户群以传单、电子邮件等形式发送广告信息；那么对于商家来说，在减少人力物力来发送广告信息的同时又增加了客户看见广告后确实购买产品的反应率；对于客户来说，则会减少处理无用信息和垃圾邮件的时间．大多数基于分类的数据挖掘方法（如决策树方法[1]）都已经用来解决识别具有潜在市场价值的客户群．市场值函数挖掘方法[2,3]不同于分类方法，它可依据每个客户的市场值函数值来对客户进行排序，从而只对排在前面的顾客发送广告信息．Boosting 算法[4]是一种提高分类精度的算法，它已经成功地应用到朴素贝叶斯方法[5]、决策树方法[6]中．作者将 Boosting 算法应用到市场值函数挖掘方法中，从而改进市场值函数挖掘方法的预测效果．

1 市场值函数[2]

搜集到的客户信息保存在数据库中，形成一个训练例集合．每个客户就是一个训练例，购买某种产品的客户称为正例，尚未购买此种产品的客户称为负例．每个客户由一组属性 A 表示，市场值函数则是在这些属性上的加权和，其形式为 $r(x)=\sum_{a\in A} w_a u_a(I_a(x))$．其中：

1）x 表示一个训练例，$r(x)$ 是其市场值函数值．

2）$I_a(x)$ 是 x 在属性 a 上的取值，可用 v 表示．

3）w_a 是属性 a 的权重，表明属性的重要性，一种计算方式为

$$w_a=\sum_{v\in V_a} P(v|M)\log\frac{P(v|M)}{P(v)} \tag{1}$$

式(1)中，V_a 表示属性 a 所有取值的集合；M 表示正例的集合；$P(v)$ 表示在所有训练例中属性 a 上 v 值出现的概率；$P(v|M)$ 表示在正例集合中属性 a 上 v 值出现的概率．

4）$u_a(v)$ 是属性 a 取值为 v 时的效用函数，表明了当属性 a 取值 v 时，对 x 的市场值贡献的大小．一种计算方式为

$$u_a(v)=P(v|M)/P(v) \tag{2}$$

收稿日期：2003-06-02．
基金项目：国家自然科学基金资助项目(60173014)；北京市自然科学基金资助项目(4022003)．
作者简介：刘椿年(1944－)，男，江苏连云港人，教授，博士生导师．

2 Boosting 算法[4]

Boosting 算法的输入为一组训练例 X 和一个类别 C，Boosting 算法描述如下：

1) 初始赋予每个样本 x 相等的权重 $1/N$，即 $q(x)=1/N$，N 为训练例总数.

2) For $t=1,2,\cdots,T$ DO

① 由 $q_t(x)$ 得到一个假设 $h_t(x): X\to\mathbb{R}$

② 更新例子 x 的权重

$$q_{t+1}(x)=q_t(x)\,e^{-\beta_t y_t h_t(x)}/z_t \tag{3}$$

其中 $z_t=\sum_{x\in X} q_t(x)\,e^{-\beta_t y_t h_t(x)}$，它是一个标准化因子.

3) 输出

$$H(x)=1/\left(1+e^{\sum_{t=1}^{T}\beta_t h_t(x)}\right). \tag{4}$$

关键问题是 β_t 的选择，当 $h_t(x)$ 在 $[-1,+1]$，$\beta_t=\{\ln[(1+r_t)/(1-r_t)]\}/2$. 其中，

$$r_t=\sum_x q_t(x)\,y_x h_t(x),\quad y_x=\begin{cases}1 & x\text{属于类别 }C\\ -1 & x\text{不属于类别 }C\end{cases} \tag{5}$$

可见在 Boosting 算法中，每个例子被赋予一个权值. 在学习假设 h_t 之后，增加由 h_t 导致错误的训练例子的权值，并且通过重新对训练例子计算市场函数值，来学习下一个分类器 h_{t+1}. 这个过程重复 T 次. 最终的假设从这一系列的分类器中综合得出.

3 市场值函数算法的修改

1) 关于式(1)、(2)中 $P(v)$、$P(v|M)$ 的计算

在未应用 Boosting 算法的情况下，$P(v)=|m(v)|/|U|$. 其中：$m(v)$ 为训练例中在属性 a 上值为 v 的训练例集合，$|m(v)|$ 为其个数；U 为训练例集合，$|U|$ 为其个数.

$P(v|M)=|m(v/M)|/|M|$. 其中：$m(v/M)$ 为正例集合中在属性 a 上值为 v 的训练例集合，$|m(v/M)|$ 为其个数；$|M|$ 为正例个数.

在应用 Boosting 算法的情况下，

$$P(v)=\sum_{x\in m(v)} q(x)/\sum_{x\in U} q(x),\quad P(v|M)=\sum_{x\in m(v/M)} q(x)/\sum_{x\in M} q(x)$$

其中 $q(x)$ 为 Boosting 算法中训练例 x 的权重. 通过调整权值，体现了提升的思想.

2) 关于训练例 $h_t(x)$ 的计算

因训练例中绝大多数例子为负例，为了保证 Boosting 算法中的 $\beta_t>0$，需要把市场函数值限定在 $[-1,+1]$，对式(4)～(6)中的 $h_t(x)$ 做以下计算

$$h_t(x)=\begin{cases}r(x)/\sum_{c\in U} r(c) & r(x)\geqslant L\\ -1/e^{-r(x)} & r(x)<L\end{cases} \tag{6}$$

其中 L 为所设定的阈值.

式(6)表示，如果训练例 x 的市场函数值 $r(x)$ 大于等于所设定的阈值 L，$h_t(x)$ 修改为 $r(x)$ 和所有训练例的市场函数值之和 $\sum_{c\in u} r(c)$ 的比，使 $h_t(x)$ 限定在 $[0,1]$；如果训练例 x 的市场函数值 $r(x)$ 小于所设定的阈值 L，$h_t(x)$ 修改为 $-1/e^{-r(x)}$，使 $h_t(x)$ 限定在 $[-1,0]$.

做此修改的理由为：计算训练例的市场函数值，目的是计算下一轮各例子的权重，此时所参考的实际是各例子在所有例子中的排序，而不是它们具体市场函数值的大小．所以经过这样的修改后，既能实现把函数值限定在[－1，＋1]，满足所采用的 Boosting 算法的要求，又不会改变训练例市场函数值的次序．

4　算法的评价

本试验采用 lift 值评价方法[7]：在把测试例的市场函数值按降序排列之后，平均分成 10 份．如果算法有效，排在前面的组中会比排在后面的组出现更多的正例．

具体的计算方法有两种，一种是通过分析各组中正例个数来判断算法的好坏，另一种是采用各组中正例个数的权值和 S_{lift} 来进行评估．本实验中采用的是后一种方法．这是因为 S_{lift} 有着很好的特性：如果正例的分布是随机的(即算法无效)，S_{lift} 是 55%(如果分的份数更多，将会下降到 50%)．最好的情况是 $S_1 = \sum_{i=1}^{10} s_i$，此时 $S_{\text{lift}} = 100\%$；最坏的情况是 $S_{10} = \sum_{i=1}^{10} s_i$(其他的 $S_i = 0$)，此时 $S_{\text{lift}} = 10\%$．由此可以看出：如果算法找到了规律，$S_{\text{lift}} > 50\%$，否则 $S_{\text{lift}} < 50\%$．并且 S_{lift} 值越大，说明算法的预测效果越好．

计算过程包括：1) 把最后所得测试例的市场函数值按降序排列．2) 把测试例平均分成 10 组，对每组赋予不同的权值．计算公式为 $S_{\text{lift}} = [(1 \times S_1 + 0.9 \times S_2 + \cdots + 0.1 \times S_{10}) / \sum_i S_i] \times 100\%$．其中：$S_i$ 为每一组中正例的个数；1，0.9，…，0.1 为每一组权重，每一组权重可根据具体应用做相应修改．

5　实验结果与分析

5.1　数据集介绍

1) nec 数据集．该数据集是某个俱乐部成员信息的一个数据集．其中的 3 万例作为训练例，包括正例 3 036 例，负例 26 964 例；剩余的 2 万例作为测试例，包括正例 2 674 例，负例 17 326 例．每个例子由 29 个属性描述．

2) coil2000 数据集[8]．该数据集中的 5 822 例作为训练例，包含正例 348 例，负例 5 474 例；剩余的 4 000 例作为测试例，其中正例 238 例，负例 3 762 例．每个例子由 85 个属性来描述．

3) oradm 数据集．该数据集是 Oracle 数据挖掘测试系统中自带的一个数据集．其中的 1 399 例作为训练例，包括正例 326 例，负例 1 073 例；剩余的 1 334 例作为测试例，包括正例 320 例，负例 1 014 例．每个例子由 14 个属性描述．

5.2　实验结果及分析

本实验以 Java 为开发平台，使用 Oracle 数据库，在 3 个不同的数据集进行了测试．在实验中，公式(6)中的阈值 L 设置为：每次循环后训练例按市场函数值降序排列的次序中前 40% 分位点处例子的市场函数值．实验结果见图 1．

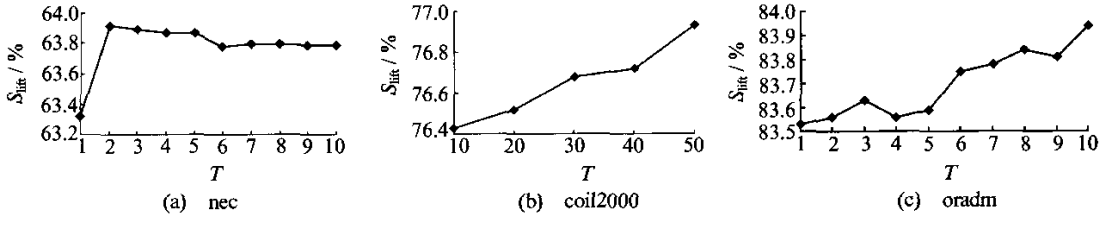

图1　不同数据集上的测试结果

Fig.1 The test results on different data sets

从实验结果可以看出：随着循环次数的增加，S_{lift} 值呈上升趋势．但需要指出的是并不是循环次数越多越好，超过了一定次数，S_{lift} 便呈下降趋势．这种现象并不奇怪，因为如果循环次数过多，训练结果就会和训练例过分吻合，此时就会出现过适应现象，从而使 S_{lift} 值下降．

6 结 论

把 Boosing 算法应用到市场函数算法中，可以提高市场函数算法的预测效果，并且此方法是切实可行的．随着 direct marketing[7] 在商业领域的应用，此方法将会有更广阔的应用前景．

参考文献：

[1] TOM M M. Machine Learning[M]. New York: The McGraw-Hill Company, Inc, 1997.

[2] YAO Y Y, ZHONG N. Mining market value function for targeted marketing[A]. 25th Annual International Computer Software and Applications Conference (COMPSAC'01) [C]. Chicago: IEEE Computer Society Press, 2001. 517-522.

[3] 黄佳进，刘椿年，李文斌. 市场值函数挖掘的研究和实现[J]. 北京工业大学学报，2003，29(1)：94-97.
HUANG Jia-jin, LIU Chun-nian, LI Wen-bin. Research and implemenitation of mining marker value function[J]. Journal of Beijing University of Technology, 2003, 29(1): 94-97. (in Chinese)

[4] ROBERT E S, YORAM S. Improved boosting algorithms using confidence-related predictions[J]. Machine Learning, 1999, 37(3): 297-336.

[5] ELKAN C. Boosting and Naive Bayesian Learning[R]. San Diego: University of California, 1997.

[6] DRUCKER H, CORTES C. Boosting decision trees[J]. Advances in Neural Information Processing Systems, 1996, 8: 479-485.

[7] LING C X, Li C H. Data mining for direct marketing: Problems and solutions[A]. Proceeding 4th International Conference on Knowledge Discovery and Data Mining[C]. New York: AAAI press, 1996. 73-79.

[8] KAYMAK U, SETNES M. Target Selection Based on Fuzzy Clustering: A Volume Prototype Approach to Coil Challenge 2000[R]. Leiden: Amsterdam and Leiden Institute of Advanced Computer Science, 2000. 1-6.

Market Value Algorithm Based on Boosting and Its Evaluation

LIU Chun-nian, CHANG Cai-qing, HUANG Jia-jin, OU Chuang-xin

(College of Computer Science of Multimedia and Intelligent Software Technology Beijing Municipal Key Laboratory, Beijing University of Technology, Beijing 100022, China)

Abstract: Identification of customers with potential market value is a key problem in direct marketing. Although some standard data mining methods may be applied for the purpose of direct marketing, no ideal results have been achieved. Therefore, based on the informational theory, the market value algorithm, is adopted to propose a linear market value function so as to discover the customers with the most potential market value. The experiment result shows that its evaluation value can be up to more than 80% and this algorithm is well interpretable. The authors try to improve the predictive ability of market value algorithm by applying Boosting to it. Experiments are carried out on three different data sets and the result shows that one percent point of evaluation value can be improved on average.

Key words: Boosting algorithm; market value function; lift value