# 一种基于用户行为特征选择的点击欺诈检测方法

董亚楠 刘学军 李 斌
(南京工业大学计算机科学与技术学院 南京 211816)

**摘 要** 在线广告是目前众多网络巨头收入的主要来源,在线广告也为网络的健康发展提供了强大的经济支撑。目前,利用用户行为属性特征来识别点击欺诈的方法中,含有较多的冗余特征,检测效率相对较低。针对这一问题,提出了一种属性特征选择与分类方法相结合的欺诈检测方法。通过训练数据集找到欺诈用户点击广告的属性特征集合,采用Fisher分方法得到了属性特征重要度排序,选取重要属性特征,并基于这些重要的特征使用支持向量机二分类方法分类。在真实数据集上的实验结果证明了该方法的可行性与有效性。

**关键词** 点击欺诈,Fisher分,支持向量机,特征选择
**中图法分类号** TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.10.027

## Click Fraud Detection Method Based on User Behavior Feature Selection

DONG Ya-nan LIU Xue-jun LI Bin
(College of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

**Abstract** Online advertisement is not only the main sources of income of profit for internet giants, but also provides powerful economic support for the internet development. The commonly used methods of click fraud detection, which are based on the features of client's behavior, may lead to inefficiency in fraud detection due to redundant features. To solve this problem, a fraud detection method which combines feature selection with classification method was proposed. According to the feature attributes set of fraud advertisement which is found through training set, attribute significance is sorted by Fisher score method. The important attributes is selected and the SVM algorithm is lastly introduced into classification based on these important attributes. Experiments on real data set demonstrate that the proposed detection method is feasible and valid.

**Keywords** Click fraud, Fisher score, Support vector machine, Feature selection

广告行业中,在线广告的投入相对于其他领域增长更快。根据美国eMarketerInc研究,早在2010年,仅美国在在线广告上面的投入就超过了100亿美元。大多数从事在线广告研究的学者认为,在基于PPC模式的在线广告系统中,大约10%~15%的广告点击并非真实有效,我们将这部分被组织起来进行点击广告、消耗广告主的广告投入来增加广告发布商收入的点击活动叫做点击欺诈。

点击欺诈为网络生态系统的健康发展带来了挑战,不仅损害了广告主的利益,同时也破坏了广告主和广告商之间的信任。因此,行之有效的欺诈点击检测方法是至关重要的。

目前,针对点击欺诈用户检测多采用用户行为特征选择方法。然而,抽取用户点击日志的全部特征用于检测,将导致数据维数过高,且各个特征项之间关系复杂。而冗余特征项的存在,不仅会增加存储代价,而且会降低检测系统性能。基于SVM技术的欺诈检测问题通常转化为一个二次规划问题来求解,但二次规划的计算量随着变量的增加而呈指数增加。对于在欺诈检测中经常遇到的高维、大规模数据的模式分类问题,如何提高基于SVM进行数据处理的实时性、缩短训练和检测时间,是当前需要解决的一个重要问题。为了适应实时异常检测的要求,有必要在SVM训练之前进行样本属性特征的选择和特征降维,以降低SVM分类器的复杂度,提高检测速度。目前,用于特征项选择的算法有粗糙集法、信息增益、支持向量机和遗传算法等。前两种算法应用于过滤器模式,速度快但选择效果略差;后两种应用于封装器模式,选择效果好,但计算复杂度却较高。比较流行的特征降维方法有主成分分析(PCA)、多维缩放(MDS)、线性判别分析(LDA)等[1]。PCA方法理论完善、概念简单,但其对于高维数据特征向量的计算较为复杂,且难以确定合适的特征维度。MDS方法能够较好地保持数据间的差异性,但无法判别所得到的主特征向量与分类标号的相关程度。LDA方法则是一种基于监督的降维方法,但它利用有类别标号的数据,当训练数据较少时会存在过拟合的现象,且对高维数据较敏感。

针对上述问题,本文提出了一种基于用户行为特征选择的欺诈检测方法,使用了与分类器无关的评价函数(Fisher分)进行特征子集选择,并使用分类器算法(SVM)评估选择结果。Fisher分计算过程较简单,且通过Fisher分值能较直

到稿日期:2015-09-02 返修日期:2015-11-29 本文受国家自然科学基金(61203072),江苏省重点研发计划(社会发展)(BE2015697)资助。
董亚楠(1991—),女,硕士生,主要研究方向为数据挖掘、计算广告,E-mail:Jud_y_n@163.com;刘学军(1971—),男,博士,教授,主要研究方向为数据库、数据挖掘、传感器网络;李 斌(1979—),男,硕士,讲师,主要研究方向为数据库、传感器网络。

观地观测出特征与类别的相关性。首先采用 Fisher 分依次计算用户行为特征，对用户行为特征的重要度排序，Fisher 分值高的特征表明了其对分类结果的影响比较大，与分类相关性也较高；进一步地选取出重要属性特征，压缩数据空间，在不损坏有效信息的前提下降低了数据处理的规模；然后基于选取的重要属性特征，采用 SVM 二分类算法处理用户点击数据，最终判断点击用户的合法性。基于 BuzzCity 的真实数据集上的实验验证了该方法在检测欺诈点击用户方面的准确性和有效性。

## 1 相关工作

目前，国内外采用的常规的防作弊方法有 IP 防止作弊、COOKIES 防止作弊、点击率阈值设置、结合 ALEXA 数据防止作弊、来源统计防止作弊、时间顺差防作弊、鼠标值、利用图形验证码等。其中，文献[2]提出了一种采用插入的虚假广告统计反馈数据的方式进而以第三方的身份帮助广告主判断来访者是机器还是人。Tuzhilin[3]研究了 Google、Yahoo 的在线过滤技术，其对点击欺诈的识别其有一定作用。但这些过滤技术都比较简单，对于一些复杂的“点击欺诈”的袭击就无能为力了。用户点击广告的行为特征可以将欺诈用户暴露出来[4]，基于用户行为的欺诈检测技术也已经被应用[5,6]。Perera 等[5]提出一个基于派生的行为属性的检测方法来提高检测欺诈点击行为效果。Immorlica 等[6]将基于广告用户的点击行为的学习算法应用于检测欺诈点击，不过此算法是针对只存有某个广告位的情况而提出的。

正常用户和欺诈用户在点击广告时的一些行为特征有着明显区别，因此可以将这些行为特征作为区分欺诈用户的条件。针对用户行为特征采用的分类方法包括神经网络、贝叶斯模型、支持向量机等。特征选择是从一组特征中去除无关的、冗余的信息，选择出一组最优特征，以降低特征空间维数的过程。特征选择在机器学习和模式识别中具有重要的作用，直接关系到学习机和分类器的效率和性能。文献[7]提出的欺诈检测特征选择方法说明经过特征选择的欺诈检测系统可以有效检测点击欺诈用户。基于分类方法的检测技术已经应用于欺诈检测技术中，且也取得了较好的检测效果[8]。

点击欺诈的行为是指以某种金钱或者商业目的为出发点，对网络广告进行恶意点击并达到消耗广告费用和抬高成本的一种行为。简单来说，当网络发布商点击其网页上的广告提高他们的收入，或企业点击竞争对手的广告来蚕食对方的广告预算时，就构成了点击欺诈。网络广告收入是当今世界各国基于互联网企业的主要收入来源，点击欺诈损害了虚拟世界的诚信基石及互联网发展的经济基石。

图 1 示出了网络广告三角色关系及流程。

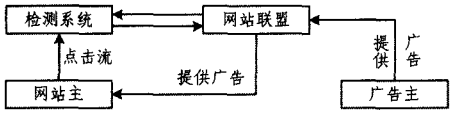


图 1 网络广告三角色关系及流程

## 2 欺诈检测

### 2.1 检测体系概述

将 Fisher 分和支持向量机(SVM)分类算法应用于检测算法中，图 2 给出了欺诈检测算法的流程。

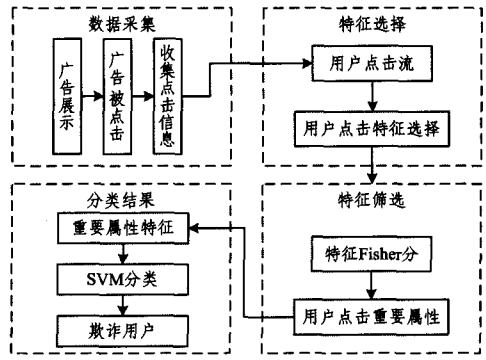


图 2 欺诈检测算法的流程

### 2.2 Fisher 分

Fisher 分是一种基于样本距离的特征选择算法[9]。当某个特征能使不同类样本之间具有最大距离，而同类样本之间具有最小距离时，Fisher 分算法赋予该特征最高的 Fisher 分值，Fisher 分值越高，表示该特征对分类的影响也越大，与分类具有较高的相关性。因此，本文使用特征的 Fisher 分值的高低代表特征与分类的相关性程度。

对于二分类问题，考虑训练集合 $X=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$x_i\in R^d(i=1,2,\cdots,N)$，$d$ 为原始特征空间的维数，类标记为 $y_i\in\{+1,-1\}$，$N$ 为训练样本数。将 $X$ 中正类样本集合记为 $X_1$，负类样本集合记为 $X_2$，$N1$ 为正类样本数，$N2$ 为负类样本数，Fisher 分值定义为：

$$F=D_b/D_i \tag{1}$$

其中，$D_b$ 为类间的离散度，描述两类样本间的距离；$D_i$ 为类内的离散度，描述同类样本间的距离。定义 $D_b=(\overline{m_1}-\overline{m})^2+(\overline{m_2}-\overline{m})^2$，$\overline{m_1}$、$\overline{m_2}$、$\overline{m}$分别为正、负类样本和所有样本的均值，$\overline{m_1}=\frac{1}{N_1}*\sum_{x\in X_1}x$，$\overline{m_2}=\frac{1}{N_2}*\sum_{x\in X_2}x$，$\overline{m}=\frac{1}{N}*\sum_{x\in X}x$；$D_i=D_1+D_2$，$D_1=\frac{1}{N_1}*\sum_{x\in X_1}(x-\overline{m_1})^2=\sigma_1^2$，$D_2=\frac{1}{N_2}*\sum_{x\in X_2}(x-\overline{m_2})^2=\sigma_2^2$，$\sigma_1^2$ 和 $\sigma_2^2$ 分别是正类和负类样本的方差。因此，$F$ 可以写为：

$$F=\frac{(\overline{m_1}-\overline{m})^2+(\overline{m_2}-\overline{m})^2}{\frac{1}{N_1}*\sum_{x\in X_1}(x-\overline{m_1})^2+\frac{1}{N_2}*\sum_{x\in X_2}(x-\overline{m_2})^2} \tag{2}$$

第 $r$ 个特征的 Fisher 分值为：

$$F_r=\frac{D_b}{D_i}=\sum_{i=1}^{2}(\overline{m_{i,r}}-\overline{m_r})^2/\sum_{i=1}^{2}\sigma_{i,r}^2 \tag{3}$$

其中，$\overline{m_{i,r}}$、$\overline{m_r}$分别为第 $i$ 类样本和所有样本的第 $r$ 个特征的均值，$\sigma_{i,r}^2$为第 $i$ 类样本第 $r$ 个特征的方差。每个特征的Fisher分值表示该特征对分类的贡献大小，即特征的重要性程度。因此，分值高的特征将被用来构建新的特征子集。这种基于 Fisher 分值的特征选择方法对于分类而言是比较简单、有效的。

### 2.3 SVM 分类算法

支持向量机(Support Vector Machine，SVM)[10]是建立在统计学习理论基础上的一种机器学习方法，经常用于各种分类和预测，它能使错误的检测率减小到最小，同时具有很好的泛化能力。SVM 是一种不太容易过拟合的分类方法，比较适合二分类问题，在二分类问题上更健壮。将 SVM 用于欺诈检测系统[11]，实验证明其能较好地检测欺诈行为。本文选择 SVM 分类算法作出最终决策。

对于上节所讨论的 $N$ 个训练样本：$\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$x_i\in R^d(i=1,2,\cdots,N)$，假设样本是线性可分的，根据统计学学习理论，为了使泛化误差最小，需要按照最大间隔原则求出最优分类超平面。SVM 分类算法将求解分类超平面看成求解一个二次规划问题[12]。

分类面两边的样本对应两种不同的类别，其分类面的线性判别函数一般形式为：

$$f(x,\omega)=\text{sign}(\omega * x+b) \tag{4}$$

在已知方向 $\omega$ 的情况下构造分类面（直线），使分类面到两边正确分类点的距离为 $2/\|\omega\|$，即相应的几何间隔为 $2/\|\omega\|$，用极大化间隔的方法可得出求解最小化 $\|\omega\|$ 的二次函数问题，即：

$$\min f(\omega)=\frac{1}{2} * \|\omega\|^2 \tag{5}$$

满足约束条件：

$$y_i(\omega * x_i+b)-1\geqslant 0,i=1,2,\cdots,N \tag{6}$$

这是一个典型的二次规划问题，经过变换，问题(5)转换为其对偶问题，即：

$$\max_{\alpha} W(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2} * \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i * x_j) \tag{7}$$

满足约束条件：

$$\sum_{i=1}^{N}y_i\alpha_i=0,\alpha_i\geqslant 0 \tag{8}$$

求解得到最优决策函数：

$$f(x)=\text{sgn}(\omega * x+b)=\text{sgn}(\sum_{i=1}^{N}\alpha_i * y_i(x_i * x)+b^*) \tag{9}$$

其中，sgn()为符号函数；$f(x)=1$ 表示 $x$ 为正常样本，$f(x)=-1$ 表示 $x$ 为异常样本；$\alpha_i^*$ 为最优解。在非线性条件下，引入 Mercer 核函数 $K(x_i,x)=\varphi(x_i)*\varphi(x)$ 代替两个向量的内积运算，将输入量 $x\in R^d$ 通过映射函数 $\varphi(x)$ 变换到高维特征空间 $F$，由特征空间的线性核运算形式实现原样本的非线性分类，相应的式(7)可转化为：

$$\max_{\alpha} W(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{10}$$

决策函数转换为：

$$f(x)=\text{sgn}(\omega * x+b)=\text{sgn}(\sum_{i=1}^{N}\alpha_i * y_i K(x_i,x)+b^*) \tag{11}$$

这样，输入空间的二分类问题就转换为二次规划问题，该规划问题有相应的快速解法。采用不同的满足 Mercer 条件的函数作为核函数，就可以构造实现输入空间中不同类型的非线性决策面的学习机器。

### 2.4 欺诈检测算法描述

本文将点击日志数据中非正常的点击数据看成异常数据，对于本文提出的欺诈检测系统的检测模块而言，它接收来自前一模块的数据结果并进行异常分析。当欺诈检测模块接收到数据后，首先，进行数据预处理，得到统一规范的数据；然后，利用 Fisher 分对点击数据的特征计算其与分类的相关性，由高到低依次选取若干特征，并计算若干特征的 Fisher 分的加权和，当该值占全部特征 Fisher 分的加权和比率达到某一阈值时，即得到所需的特征集合；最后，根据得到的新的特征集及相应的原始数据重新生成新的训练样本集，该样本集仅保留了影响分类效果的重要特征，对最终的训练样本进行数值化和归一化处理后，输入 SVM 训练器训练，再通过测试样本对检测模型进行测试，输出分类结果。整个欺诈检测流程描述如下。

输入：广告点击日志数据集（点击日志数据特征集）

输出：分类结果

开始

1. 数据预处理，得到统一规范的数据；
2. 计算每个特征属性的 Fisher 分值，依据 Fisher 分对特征重要性由高到低排序；
3. 预设阈值 $\eta(0<\eta<1$，$\eta$ 初始值取 0.9)；（$\eta$ 的取值影响最后的分类精度和分类效率）
4. For i=1 to n(n 为特征个数)
   {i. 计算 1 到 i 的 i 个特征的 Fisher 分的加权和 A；
   ii. 计算全部特征的 Fisher 分的加权和 B；
   iii. 若 $A/B\geqslant\eta$，n 取使得 $A/B\geqslant\eta$ 的最小值，提取出前 n 个特征集合；
   否则，继续执行步骤 i。
   随机地将数据集分为训练集 $X_{train}$ 和验证集 $X_{valid}$；
   使用 $X_{train}$ 建立 SVM 分类模型，以 $X_{valid}$ 对模型进行验证；
   计算平均验证误差 $\varepsilon$；}
   选择具有最低平均验证误差 $\varepsilon$ 时的阈值；
5. 丢弃不符合所选阈值的特征属性；
6. 基于得到的属性特征集合，将测试集输入 SVM 模型分类器，得出分类结果。

结束

## 3 实验结果与分析

本文实验采用 Singapore Management University（SMU）2012年在加拿大组织的 FDMA2012 竞赛提供的一个点击欺诈检测的标准数据集（http://palanteer.sis.smu.edu.sg/fdma2012/）。该数据集由 BuzzCity 移动广告网络公司提供。数据集包含有训练集和测试集，分不同的时间段获取得来。数据集欺诈用户与正常用户所占的比例分别为 2.298%，97.702%。训练集包括点击训练集和用户训练集，有 3173834 个实例，其中每个实例代表着一个用户的点击记录。而用户训练集有 3081 个附有标签（欺诈或者合法）的用户的记录，通过训练集数据训练得到最终分类模型来评估测试集数据，测试集数据也包括与训练集相似的数据，同样包括点击测试集 2598815 条实例和用户测试集 2112 条实例。

在数据集中，用户的一次点击即生成一条点击日志记录，记入点击日志集。用户的每条点击记录由点击行为属性构成，每次点击含有 118 个属性，完整的属性列表可以从 http://clifton.phua.googl-epages.com/feature-list.txt 中获取。部分属性列表及含义如表 1 所列。

表 1 点击欺诈属性及含义描述

| 属性名称 | 含义描述 |
|---|---|
| Id | Unique identifier of a particular click |
| Iplong | Public IP address of a clicker/visitor |
| Agent | Phone model used by a clicker/visitor |
| partnerid | Unique identifier of a publisher |
| cid | Unique identifier of a given advertisement campaign |
| Cntr | Country from which the clicker/visitor is |
| Timeat | Timestamp of a given click |
| Referrer | URL where ad banners are clicked |
| Category | category Publisher's channel type |

### 3.1 属性选择

由于原始点击日志数据集属性太多，且包含一些对分类

结果(正常与欺诈)影响很小的属性,根据经验[10],首先对 118 个属性($f_1, f_2, \cdots, f_{118}$)做筛选,即通过 2.4 节所述步骤进行属性选择实验,排除对分类结果影响很小的属性,从而得到相应分值高的属性集 $F$,即此属性集 $F$ 内的属性对分类结果影响较大。这样就将原始数据集维度降低,从而提高了检测的速率。

### 3.2 属性的数值化

经过特征属性选择后的属性集减小了原始数据集的维数。根据 3.1 节属性选择的结果,使用不同阈值 $\eta$ 进行属性选择,$\eta$ 分别取值为 0.9,0.85,0.8,0.75 时,选择其对应的属性个数 $n$,即依次选择前 16 个、前 12 个、前 8 个、前 6 个重要属性来构造 16 维特征子集 $f_1-f_{16}$、12 维特征子集 $f_1-f_{12}$、8 维特征子集 $f_1-f_8$、6 维特征子集 $f_1-f_6$。经属性选择后的结果大大缩小了点击数据记录的维数,从而有利于提高 SVM 算法的分类效率。由于 SVM 只能处理数值化向量,因此在把训练样本输入 SVM 学习器前,必须对特征属性进行预处理。这里采用了与文献[13]一样的处理方法,即引入异构值差度量(HVDM)距离函数来反映不同属性对异构数据集样本点间距离的贡献。假设异构数据集 $X$ 上两个数据 $x$,$y$ 的第 $i$ 个连续属性分别为 $x_i, y_i$,则 $x, y$ 在第 $i$ 个属性上的距离为:

$$\text{normalized } diff_i(x,y)=\frac{|x_i-y_i|}{4\sigma_i} \tag{12}$$

其中,$\sigma_i$ 为数据集上第 $i$ 个属性的方差。

假设异构数据集 $X$ 上两个数据 $x, y$ 的第 $j$ 个离散属性分别为 $x_j, y_j$,则 $x, y$ 在第 $j$ 个属性上的值差度量为:

$$\text{normalized } vdm_j(x,y)=\sqrt{\sum_{j=1}^{C}\left|\frac{N_{j,x,c}}{N_{j,x}}-\frac{N_{j,y,c}}{N_{j,y}}\right|^2} \tag{13}$$

其中,$N_{j,x}$ 表示的是数据集 $X$ 上所有数据第 $j$ 个属性取值为 $x_j$ 的数据个数,$N_{j,x,c}$ 表示的是数据集 $X$ 上所有数据第 $j$ 个属性取值为 $x_j$ 并且输出类别为 $C$ 的数据个数,$C$ 表示数据输出类别。

异构值差度量(HVDM)距离函数 $H(x,y)$ 定义为:

$$H(x,y)=\sqrt{\sum_{i=1}^{m}d_i^2(x_i,y_i)} \tag{14}$$

其中,$d_i(x_i,y_j)$ 取值如下:当 $x_i$ 或者 $y_i$ 值为空,值取 1;当 $x_i$ 或者 $y_i$ 为连续属性,由式(12)计算该值;当 $x_i$ 或者 $y_i$ 为离散属性,由式(13)计算该值。

### 3.3 SVM 分类结果

根据上述属性选择结果,在不同特征子集条件下,构造了 3 个训练样本集和 3 个测试样本集,训练样本集和测试样本集均来自原始数据集。训练样本集 1($Tr1$)由合法用户点击记录组成,训练样本集 2($Tr2$)由欺诈用户点击记录组成,训练样本集 3($Tr3$)由合法、欺诈用户的混合点击记录组成,测试样本集的构成与此类似。实验中,使用 LibSVM 作为 SVM 训练和测试的工具[14],参数 $C$ 使用默认值,选用线性核 $K(x, x_i)=x*x_i$。

在特征选择算法运行过程中,最终选择的属性个数受算法中阈值 $n$ 的影响,$n$ 的取值代表了数据的维度,因此会对 SVM 分类器的性能产生影响。通过考察阈值 $n$ 的变化,即改变选取的属性维度(下文将 $n$ 统称为属性维度)来评估其对 SVM 分类器性能的影响。实验结果如表 2 所列。

表 2 不同属性维度 $n$ 的 SVM 分类器的欺诈检测结果

| 属性维度 n | 训练时间(s) | 检测时间(s) | 准确率(%) |
|---|---|---|---|
| 16 | 11.23 | 5.49 | 89.9 |
| 12 | 9.34 | 4.35 | 89.88 |
| 8 | 5.53 | 2.02 | 89.86 |
| 6 | 5.4 | 1.9 | 87.53 |

从表 2 可以看出,综合考虑训练时间、测试时间和准确率的情况,$n$ 取值越小,分类器的训练和检测时间越短,准确率也越低。同时可以看出,$n$ 取 8 时的 SVM 分类准确率与 $n$ 取 16,12 时的 SVM 的分类准确率相当,但训练时间和测试时间有显著降低。而 $n$ 取 6 时的训练时间和检测时间虽然降低,但 SVM 分类准确率却明显降低。

在其余参数保持不变的情况下,考察属性特征选择前后对 SVM 分类性能的影响。为保证数据的均衡性,在选择样本集数量时,不同类型的样本集数目应尽量保证均衡。通过对 SVM 分类器输入不同类型的数据进行实验,实验结果如表 3 和表 4 所列。

表 3 属性特征选择后的 SVM 欺诈检测结果

| 数据类型 | 训练时间(s) | 检测时间(s) | 检测准确率(%) |
|---|---|---|---|
| 合法(Tr1) | 5.61 | 2.17 | 89.93 |
| 欺诈(Tr2) | 4.11 | 1.73 | 89.95 |
| 混合(Tr3) | 6.12 | 2.6 | 89.89 |

表 4 全部属性特征的 SVM 欺诈检测结果

| 数据类型 | 训练时间(s) | 检测时间(s) | 检测准确率(%) |
|---|---|---|---|
| 合法(Tr1) | 6.31 | 2.4 | 89.95 |
| 欺诈(Tr2) | 4.76 | 1.93 | 89.98 |
| 混合(Tr3) | 10.73 | 4.35 | 89.9 |

由表 3、表 4 的结果可知,经过属性特征选择后的 SVM 欺诈检测与未经属性特征选择的 SVM 欺诈检测结果相比,两者在检测准确率上结果相当,而前者在训练时间和检测时间上有所缩短,即该方法的检测效率有所提高。可见,冗余属性的存在不但不能提高分类器检测性能,还会增加分类器的负担。

对于检测点击欺诈方法而言,希望检测方法在检测欺诈方面能够获得较高的精度。本文使用 ROC 曲线来评估所提方法。ROC 曲线是以真阳性率为纵坐标,假阳性率为横坐标绘制的曲线。真阳率(精度)定义为 $TPR=\frac{TP}{TP+FP}$,假阳率(误检率)定义为 $FPR=\frac{FP}{TN+FP}$,$TP$ 表示正类判定为正类的数量,$FP$ 表示负类判定为正类的数量。本文提出的方法中,不同维度属性集建立的 SVM 分类器的 ROC 曲线如图 3 所示。

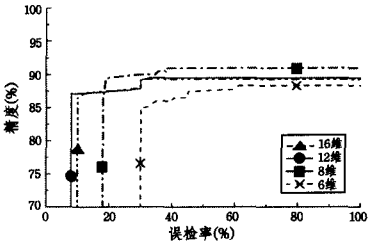


图 3 不同维度属性集建立的 SVM 分类器的 ROC

由图 3 可以看出,当误检率在 10%左右时,相同精度下,16 维、12 维属性集误检率略低。当误检率超过 15%时,8 维属性集具有更高的检测精度。当属性维度取 6 时的误检率和

精度的取值都不理想。ROC曲线下的面积AUC的值越接近1,表明分类器的分类效果越好。从图3可以看出,8维属性集的AUC值最高。综合考虑训练时间和检测时间,选取8维属性集作为分类器的属性集取得了较好的分类效果。

**结束语** 特征选择问题是欺诈检测的核心问题之一,而有效的特征提取对于提高欺诈检测的检测率、降低欺诈检测的误报率、提高欺诈检测的实时性有着重要影响。文中提出的检测方法首先对用户点击广告的日志数据特征利用Fisher分度量其特征重要性,然后选择相对重要的属性利用SVM二分类方法对用户最终分类。文中基于用户行为的欺诈点击检测算法经过了真实数据集训练,实验证明,通过去除冗余的特征属性,采用最能反映欺诈用户行为的重要特征进行检测欺诈,分类器的精度与未去除冗余特征所构建的分类器精度相当,但是训练和测试时间有所降低,且能够保证在精度基本不改变的情况下,更快速和有效地检测欺诈用户。

## 参 考 文 献


[1] Chen Shi-guo, Zhang Dao-qiang. Experimental Comparisons of Semi-Supervised Dimensional Reduction Methods[J]. Journal of Software, 2011, 22(1): 28-43(in Chinese)
陈诗国,张道强. 半监督降维方法的实验比较[J]. 软件学报, 2011, 22(1): 28-43

[2] Haddadi H. Fighting online click-fraud using bluff ads[J]. ACM SIGCOMM Computer Communication Review, 2010, 40(2): 21-25

[3] Tuzhilin A. The Lane's Gifts v. Google Report[EB/OL]. 2006 [2013-03-01]. http://googleblog. blogspot. com/pdf/Tuzh-ilin-Report. pdf

[4] Qin Chao. Visitor action analyzing system for electronic business website[D]. Shanghai: Shanghai Jiao Tong University, 2006(in Chinese)
秦超. 电子商务网站访客行为分析系统[D]. 上海:上海交通大学, 2006

[5] Perera K S, Neupane B, Faisal M A, et al. A Novel Ensemble Learning-Based Approach for Click Fraud Detection in Mobile Advertising[M]// Mining Intelligence and Knowledge Exploration. Springer International Publishing, 2013: 370-382

[6] Immorlica N, Jain K, Mahdian M, et al. Click Fraud Resistant Methods for Learning Click-Through Rates[C]// Proceedings of the Workshop on Internet and Network Economics. Berlin Heidelberg: Springer, 2005: 34-45

[7] Oentaryo R, Lim E P, Finegold M, et al. Detecting Click Fraud in Online Advertising: A Data Mining Approach[J]. Journal of Machine Learning Research, 2014, 14(1): 99-140

[8] Hager M, Landergren T. Implementing best practices for fraud detection on an online advertising platform [D]. Gothenburg: Chalmers University of Technology, 2010

[9] Sergios T, Konstantinos K. Pattern recognition(2nd ed)[M]. Salt Lake City: Elsevier Academic Press, 1999

[10] Chang C C, Lin C J. LIBSVM: A library for support vector machines[C] // ACM Transactions on Intelligent Systems and Technology. 2011: 389-396

[11] Ravisankar P, Ravi V, Raghava Rao G, et al. Detection of financial statement fraud and feature selection using data mining techniques[J]. Decision Support Systems, 2011, 50(2): 491-500

[12] Cortes C, Vapnik V. Support vector networks [J]. Machine Learning, 1995, 20(3): 273-297

[13] Zhang Yi-rong, Xian Ming, Xiao Shun-ping, et al. An Anomaly Intrusion Detection Technique of Support Vector Machine Based on Rough Set Attribute Reduction[J]. Computer Science, 2006, 33(6): 64-68(in Chinese)
张义荣,鲜明,肖顺平,等. 一种基于粗糙集属性约简的支持向量异常入侵检测方法[J]. 计算机科学, 2006, 33(6): 64-68

[14] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent & Technology, 2011, 2(3): 389-396


(上接第134页)


[2] Dantu R, Loper K, Kolan P. Risk management using behavior based attack graphs[C]// International Conference on Information Technology: Coding and Computing(ITCC 2004). Las Vegas: IEEE, 2004: 445-449

[3] Poolsappasit N, Dewri R, Ray I. Dynamic security risk management using bayesian attack graphs[J]. IEEE Transactions on Dependable and Secure Computing, 2012, 9(1): 61-74

[4] Albanese M, Jajodia S, Noel S. Time-efficient and cost-effective network hardening using attack graphs[C]// 2012 42nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, 2012: 1-12

[5] Wang S, Zhang Z, Kadobayashi Y. Exploring attack graph for cost-benefit security hardening: A probabilistic approach[J]. Computers & Security, 2013, 32: 158-169

[6] Luo Zhi-yong, Sun Guang-lu, Liu Jia-hui, et al. Application of attack graphs algorithms in intrution prevention system[J]. Journal of Yunnan University, 2012, 34(3): 271-275(in Chinese)
罗智勇,孙广路,刘嘉辉,等. 攻击图算法在入侵防御系统中的应用[J]. 云南大学学报(自然科学版), 2012, 34(3): 271-275

[7] Liu Gang, Li Qian-mu, Zhang Hong. Defense strategy generation method for network security based on state attack-defense graph [J]. Journal of Computer Applications, 2013, 33(S1): 121-125 (in Chinese)
刘刚,李千目,张宏. 基于状态攻防图模型的网络安全防御策略生成方法[J]. 计算机应用, 2013, 33(S1): 121-125

[8] Mell P, Scarfone K. Improving the common vulnerability scoring system[J]. IET Information Security, 2007, 1(3): 119-127

[9] Wang Yu-long, Yi Yang. PVL: A Novel Metric for Single Vulnerability Rating and Its Application in IMS[J]. Journal of Computational Information Systems, 2012, 8(2): 579-590

[10] Spanos G, Angelis L. Impact Metrics of Security Vulnerabilities: Analysis and Weighing [J]. Information Security Journal A Global Perspective, 2015, 24(1-3): 1-15

[11] Ye Yun, Xu Xi-shan, Jia Yan, et al. An Attack Graph Based Probabilistics Computing Approach of Network Security[J]. Chinese Journal of Computers, 2010, 33(10): 1987-1996(in Chinese)
叶云,徐锡山,贾焰,等. 基于攻击图的网络安全概率计算方法[J]. 计算机学报, 2010, 33(10): 1987-1996

[12] Li Qing-peng, Wang Bu-hong, Wang Xiao-dong, et al. Network security assessment based on probabilities of attack graph nodes [J]. Application Research of Computers, 2013, 30(3): 906-908 (in Chinese)
李庆朋,王布宏,王晓东,等. 基于攻击图节点概率的网络安全度量方法[J]. 计算机应用研究, 2013, 30(3): 906-908