支持向量机在分析型 CRM 中的应用研究

杨启仁[1],杜圣东[2]

(1.贵州民族学院 计算机与信息工程学院,贵州 贵阳 550025;
2.西南交通大学 CAD 工程中心,四川 成都 610031)

摘 要:在支持向量机分类模型的基础上,以客户流失预测为例,阐述了分析型(CRM)体系结构和客户主题数据集市的设计,并详细介绍了数据预处理、模型创建及评估的方法步骤。通过对移动运营商 CRM 系统中的客户数据进行实证研究表明,把支持向量机应用于分析型 CRM 中的客户流失挖掘是有效可行的。

关键词:支持向量机;数据挖掘;分析型 CRM;客户流失模型

中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2010)13-0072-04

Application of support vector machine in analytic CRM

YANG Qi Ren[1], DU Sheng Dong[2]

(1.School of Computer and Information Engineering, Guizhou University for Nationalities, Guiyang 550025, China;
2.CAD Engineering Center, Southwest Jiaotong University, Chengdou 610031, China)

Abstract: Based on support vector machine classification model, this paper takes prediction of customer losing for example, and explains the architecture of the analytic CRM and the design of the customer data mart. Then introduces the methods of data pretreatment, model building and model assessment in detail. Through the experiment which based on the real customer data of mobile operators's CRM, the results show that application of support vector machine in mining of analytic CRM on prediction of customer losing is efficient and feasible.

Key words: support vector machine; data mining; analytic CRM; prediction model of customer losing

随着通信市场竞争的加剧,移动运营商之间对客户的争夺也日趋激烈。各运营商都有自己完整的运营支撑系统,如计费系统、帐务系统、营业系统和客户服务系统等。这些系统累积了海量的客户相关数据,很多企业也都拥有自己的客户关系管理 CRM(Custom Relationship Management)系统[1]。如何通过数据挖掘技术对 CRM 系统中累积的大量历史数据进行分析处理,以提供有效的决策知识,从而获得新客户,提高客户满意度、防止客户流失是分析型 CRM 的目标。分析型 CRM[2](Analytic CRM)是创新和使用客户知识(在这一过程中采用数据仓库、OLAP 和数据挖掘技术对客户数据进行分析,提炼出有用信息),帮助企业提高优化客户关系的决策能力和整体运营能力的概念、方法、过程以及软件的集合。CRM 从上世纪 90 年代初基于部门级的专用解决方案,(如销售队伍自动化、客户服务支持)发展到现在以客户为中心的整体解决方案,尤其是 Internet 的迅猛发展与成熟的电子商务平台,大大推进了应用的广度和深度。目前,数据挖掘与 CRM 相结合的分析型 CRM 相关技术的研究与应用成为学术界和工业界研究的热点。

统计学习理论[3]是一种专门研究小样本情况下机器学习规律的理论,支持向量机 SVM(Support Vector Machine)作为一种新的数据挖掘技术,是在统计学习理论的基础上发展起来的新的学习算法。由于其基于结构风险最小化原则,即由有限的训练样本集得到较小的误差以确保对独立的测试样本集仍保持较小的误差,因此能有效地解决过学习问题,具有良好的推广性;另外,由于 SVM 算法能解决凸优化问题,局部最优解就是其全局最优解,因此具有较好的分类准确性。这些优良特性使得 SVM 成为继人工神经网络 ANN(Artifical Neural Network)[4]和模式识别之后的又一研究热点。最有代表性的

是美国邮政手写数字库识别研究成功地应用了SVM。在其他应用领域,如人脸识别、语音识别、模式识别、图像处理及文本分类等方面也取得了大量的研究成果。

本文在研究支持向量机并将其应用于分析型CRM的过程中,以移动通信作为分析型CRM系统的典型应用行业,其原因除了满足更激烈的商业竞争外,还在于其拥有较为完整的、规范化的并对其发展战略十分重要的客户数据基础。根据CRM中的客户历史数据对未来客户流失的可能性进行预测评估,为决策者提供有用知识具有一定的实用意义。

1 支持向量机(SVM)

VAPNIK V提出的SVM理论[5]最基本的思想之一是结构化风险最小化原则SRM(Structural Risk Minimization),该理论优于传统的经验风险最小化原则ERM(Empirical Risk Minimization)。不同于ERM试图最小化训练集上的误差的做法,SRM试图最小化VC维的上界(SRM和VC维理论见参考文献[6]),与传统的降维方法相反,SVM通过提高数据的维度把非线性分类问题转换成线性分类问题,较好地解决了传统学习算法(如人工神经网络)中训练集误差最小而测试集误差仍较大的问题,算法的效率和精度都有很大提高。近年来该方法成为构造数据挖掘分类模型和数据挖掘回归预测模型的一项新型技术。

1.1 SVM分类算法

SVM是通过构造一个最优超平面,对二值分类问题进行分割。所谓最优分类面就是要求分类面不但能将二值分类正确分开(保证经验风险最小),而且使分类间隔最大。

以对m个样本:$(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)$求解最优分类超平面为例,求解系数$w$和$b$,使超平面$(wx)+b=0$达到分类误差小、推广能力强的要求。必须满足最优分类超平面的条件:

$$y_i[(wx_i)+b]\geqslant 1,(i=1,2,\cdots,m) \tag{1}$$

$$\min_{w}\ \phi(w)=\|w\|^2 \tag{2}$$

根据最优化理论,利用Lagrange函数将其转化为求解标准型二次型规划问题:

$$\max\ W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{3}$$

$$\text{s.t.}\ \sum_{i=1}^{m}\alpha_i y_i=0,\ \ \alpha_i\geqslant 0,(i=1,2,\cdots,m) \tag{4}$$

求解上式得最优分类决策函数为:

$$f(x)=sign\{\sum_{\alpha>0}\alpha_i y_i K(x_i,x)-b_0\} \tag{5}$$

b_0可由约束条件$\alpha_i[y_i(w^{\mathrm{T}}x_i+b)-1]=0$求解,$\alpha_i$不为零的样本即为支持向量。

对于非线性二元分类,则通过某种事先选择的非线性映射(即核函数),将输入向量x映射到一个高维特征空间中,然后在这个高维空间中构造最优分类超平面,这种方法通过核函数做升维处理避免了在高维特征空间中进行复杂的运算。

1.2 SVM分类预测模型

由于现有的SVM分类模型[7]用于数据挖掘还处于试验阶段,通常只对训练好的模型做简单的测试。虽然测试模型可以对该模型的推广性能做出一些定量分析,但在现实中该分类模型是否真正实用还需了解其特点,如模型推广性、模型稳定性等。可将SVM分类模型应用于分析型CRM的客户流失分类预测,分类模型的完整建立过程分为:学习阶段、测试阶段和评估阶段。

1.2.1 学习训练阶段

(1)从客户主题数据集市中抽取客户相关数据建立训练样本集:

$$(x_1,y_1),\cdots,(x_l,y_l),x_i\in R^n,y_i\in\{-1,+1\} \tag{6}$$

(2)选择合适的核函数及核参数,作为高维特征空间在低维输入空间的一个等效形式;

(3)对输入训练样本进行规范化,将输入数据限定在核函数要求的范围之内;

(4)构造核矩阵$H(n,n)$;

(5)在式(7)约束条件下,最大化式(8),以求解拉格朗日系数a;

$$0\leqslant\alpha_i\leqslant C,\ \sum_i\alpha_i y_i=0 \tag{7}$$

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{8}$$

(6)找出支持向量SV,求解分类超平面系数b;

(7)建立训练数据的最优决策超平面,完成训练过程。

1.2.2 测试阶段

(1)装入SVM学习阶段的有关数据,包括训练数据,系数a、b,以及得到的支持向量SV;

(2)根据

$$f(x)=sign(\sum_{i=1}^{n}a_i^* y_i K(x_i,x)+b^*) \tag{9}$$

计算新输入测试数据样本的相应决策输出值;

(3)利用指示函数将$f(x)$归为$\{-1,+1\}$,做出分类决策。

1.2.3 评估阶段

在用实验数据训练和测试模型时,只是对该模型的预测效果作简单的对比,如果训练好的模型实际输出与预测输出误差很小,可认为该模型推广能力强。但现实中的数据是多变的,只是用历史数据进行预测,并不能表明该模型在后续预测中一直会有好的效果。本文所提出的评估阶段实际上是预测模型的试运行过程,在该过程中,把现实中的数据输入测试好的模型,根据输出对

模型作一些优化和调整。

以上三个阶段是一个循环往复的过程:首先用训练集建立初始模型,将测试集输入训练好的初始模型得出测试误差,如果误差较大则反复修正初始模型,当修正后的模型效果达到要求时,再用评价数据集对该模型进行评价,如果评估效果不好,则返回修正模型,如此反复直到得出最优的分类预测模型。

2 分析型 CRM

2.1 分析型 CRM 体系结构

分析型 CRM 体系结构如图 1 所示,分为数据源层、数据存储层、应用支持层和用户交互层。

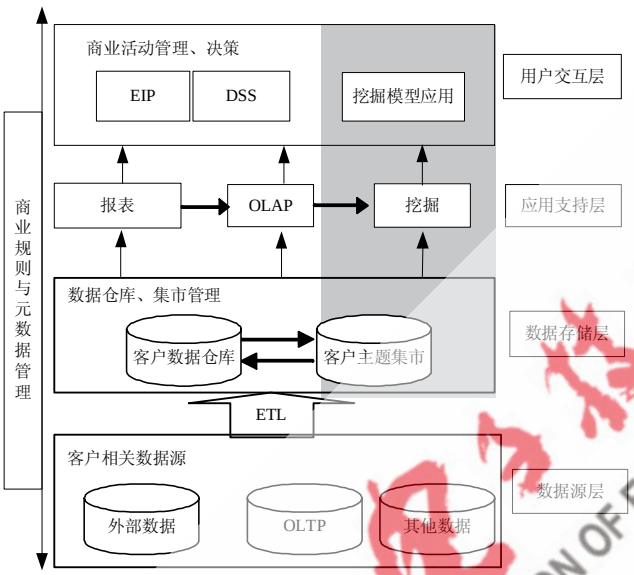

图 1 分析型 CRM 体系结构

(1)数据源层包括了企业常用信息系统和一些外部系统的数据源,如涉及客户交互的一些交易系统和服务系统,但各系统间的客户数据是分散的,而且可能重合,会出现不一致的问题。

(2)数据存储层是为了整个企业有集中统一的客户视图,通过从各源系统抽取数据,进行整合的数据仓库,在客户数据仓库的基础上,可以建立相关分析的客户主题数据集市。

(3)应用支持层除了支持复杂、智能化报表查询外,还支持 OLAP 分析,提供数据挖掘功能。

(4)用户交互层提供分析、挖掘结果,企业管理、决策层和企业其他服务人员与客户的交互形成反馈机制,从而有效地利用分析和挖掘得到有用知识。

本文研究重点是阴影板块部分:

(1)在企业已有CRM 数据仓库的基础上,抽取出客户流失预测相关的数据,建立相关主题的客户数据集市;

(2)从客户主题数据集市中抽取客户流失相关表的一些关键属性字段,形成 SVM 分类预测挖掘模型的输入数据;

(3)通过对 SVM 分类预测模型的训练和验证,并对最优模型进行应用,进一步验证反馈,形成比较稳定的客户流失分类预测模型。

2.2 分析型 CRM 主题数据集市设计

通信行业主要采用事实表和维表的形式建立数据仓库。在建立数据集市过程中重点考虑 BOSS 系统和分析型 CRM 的接口,不仅要实现物理上的转化,而且还要在逻辑上实现从 BOSS 系统实体到数据仓库实体的成功过渡。这是因为数据仓库的数据不再是业务类型,而是按主题组织。如 BOSS 系统中含有客户管理类实体、计费账务管理类实体等;而数据仓库则分为客户主题、账务主题等。

本文中客户主题数据集市[8]是从 CRM 数据仓库中抽取客户数据、业务数据、帐务数据等信息,这些数据经过转换、装载、聚合进入到接口数据层,可作为客户流失分类预测模型的基础数据;数据模型层再根据模型需求对接口层数据进行汇总,生成客户流失分类预测挖掘模型输入的宽表,总体数据集市结构如图 2 所示。

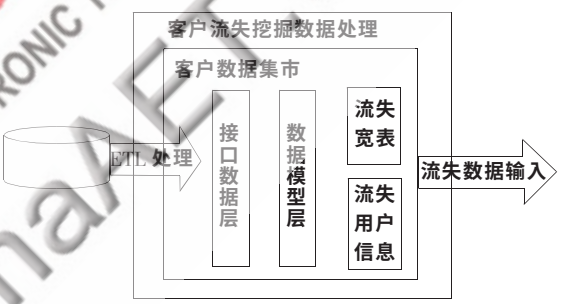

图 2 客户主题数据集市结构图

3 实证研究

3.1 电信数据处理

本文针对流失挖掘的需求建立了相关的客户主题数据集市,从客户数据仓库中抽取流失分类预测挖掘主题相关的数据,即提取与客户流失因素相关的属性,并且选择部分数据作为训练集。涉及到的数据源(这里只列出有代表性的字段,实际模型调整过程中,个别字段和属性可根据业务建议和模型本身特点添加或者删减)如表 1 所示。

表 1 数据源描述

数据源	选择属性	抽取周期
基本信息	客户 ID、年龄、性别	月
用户帐单	业务类型、月平均消费、月累积欠费	日
用户话单	月通话次数、时长、月长途次数、时长	日
客户服务	投诉业务类别、投诉次数	周

在提取的与流失因素相关的属性中,既有单粒度属性,又有多重粒度的属性,还有派生属性。在属性选择的过程中,用到了属性归约和泛化技术,最终选取表1中的属性作为模型输入字段,客户流失标记(在网、流失)作为模型输出。客户流失标记的处理如下:在2个月的预测期和1个月的评估期中,正常客户可以呈现出多种异常状态。文中以其中3种状态为流失倾向的客户特征,对其做流失标记:

(1)拆机。

(2)2个月零通话(2个月总通话次数=0且总发短信次数=0)。

(3)2个月低额消费(每个月通话次数≤5且每个月发短信次数≤5),代表一定的流失倾向。

流失分类预测模型利用3个月的的历史数据对客户在未来2个月的流失倾向进行预测,用未来第3个月的数据进行评估。本文选择基于200501~200504月之间3个月的客户数据对SVM模型进行训练,用200505~200506之间1个月的数据进行预测,用200507月的客户数据进行评估。

3.2 实验结果分析

经过数据预处理后,形成了模型输入的汇总表(即宽表),输入到本文的SVM分类预测模型中进行训练、预测和评估。模型指标评价如图3所示,模型的评价指标主要是查全率和查准率,具体指标如下:

查准率=命中用户/预测离网用户

查全率=命中用户/实际离网用户

图3 模型指标评价

通过对SVM模型的反复调整,形成最优模型时各处理阶段的数据如表2所示。

表2 实验结果

阶段	算法	查准率/%	查全率/%
训练	SVM	99.37	99.21
	ANN	99.54	99.76
测试	SVM	86.29	83.50
	ANN	81.24	70.48
评估	SVM	80.53	74.62
	ANN	79.90	53.52

从表2的实验数据可以看出,本文中的SVM分类模型相对ANN分类模型,做客户流失分类预测和评估时的查全率和查准率都有一定提高。在训练阶段,由于ANN存在过度训练情况,查全率和查准率都比SVM的训练精度要高;而测试阶段,SVM模型良好的推广性得到了验证,相比ANN的查全、查准率有较大提高;在评估阶段,SVM分类模型相对于ANN更是表现出了很好的稳定性。

分析型CRM在各领域的应用已经十分广泛,能否有效地应用数据挖掘技术对于分析型CRM十分关键。本文将支持向量机这种新的数据挖掘方法应用于移动领域客户流失挖掘,对客户离网的可能性进行预测,为决策者提供有用知识。实验中对SVM和ANN这两种模型用于流失分类预测的效果进行了对比,结果显示SVM相比ANN具有更优的分类预测效果和更好的模型稳定性,从而验证了SVM应用于分析型CRM中的客户流失挖掘是有效可行的。

参考文献

[1] BARNES J G. Secrets of customer relationship management [M]. McGraw. Hill Education, 2001.

[2] BERSON A, SMITH S, THEARLING K. Building data mining applications for crm [M]. McGraw. Hill Education, 1999.

[3] VAPNIK V. The nature of statistical learning theory [M]. New York, Springer, 1995.

[4] ALMEIDA J S. Predictive non-linear modeling of complex data by artificial neural networks [J]. Curr Opin Biotechnol. 2002,13(1):72-6.

[5] CRISTIANINI N, SHAWE J. An introduction to support vector machines, Cambridge [M]. U.K Cambridge University Press, 2000.

[6] CHERKASSKY V, SHAO X, MULIER F, et al. Model complexity control for regression using VC generalization Bounds[J]. IEEE Transaction on Neural Networks, 1999,10(5):1075-1089.

[7] 田盛丰, 黄厚宽. 基于支持向量机的数据库学习算法[J].计算机研究与发展,2000,37(1):17-22.

[8] 陈洁馨.数据仓库与决策支持系统[M].北京:科学出版社,2005.

(收稿日期:2009-12-02)

作者简介:

杨启仁,男,1973年生,硕士,讲师,主要研究方向:数据挖掘、网络安全。

杜圣东,男,1981年生,硕士,讲师,主要研究方向:数据挖掘、信息检索。