K-means与朴素贝叶斯在商务智能中的应用

赵 敏,倪志伟,刘 斌
(合肥工业大学,安徽 合肥 230009)

摘 要:不同的客户给企业带来的效益并不相同,为了提高企业的客户关系管理水平,采用基于K-means的聚类的Naive Bayesian算法来预测客户价值,从而使企业可以针对不同的客户采用不同的营销策略,为企业决策提供依据。朴素贝叶斯分类模型是一种简单有效的分类方法,它理论基础好,分类精度高,由于朴素贝叶斯分类中的独立假设前提,使得在特征选择步骤能否准确有效的分类显得尤为重要。实验结果表明,该算法能在保证一定的准确率的同时,可以预测出更多的潜在高价值客户。

关键词:客户关系管理;RFM模型;朴素贝叶斯;聚类
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2010)04-0179-04

Application Research of K-Means Clustering and Naive Bayesian Algorithm in Business Intelligence

ZHAO Min, NI Zhi-wei, LIU Bin
(Hefei University of Technology, Hefei 230009, China)

Abstract: Different customers benefits to enterprise are not the same, in order to improve the level of the enterprise customer relationship management, use the Naive Bayesian algorithm based on the K-means clustering to forecast the customer value, so that enterprises can use different marketing strategies for different customers. And this will provide a basis for business decisions. Naive Bayesian Classification Model is a simple but efficient solution, and it has solid theory foundation and high accuracy rate of classification, an effective feature selection is very important for an NB-based classifier which uses the conditional independence assumption. Experimental results show that the algorithm can guarantee a certain degree of accuracy and can predict more high-value potential customers.

Key words: CRM; RFM model; Naive Bayesian; clustering

0 引 言

近年来,销售行业的竞争越来越激烈,各个商家不仅在产品品质上精益求精,而且在企业管理信息化和销售上也做出了巨大的改进,开始从以产品为中心转移到以客户为中心的策略。随着商品销售数量的增加,大量的客户的产生,对于销售部门来说,这些客户的重要性是不同的,在实际应用中,由于不同级别不同数量的客户对于企业带来的价值不同,同时受维持客户关系费用开销的限制,企业只能针对比较重要的客户开展活动,这样给客户价值预测问题提出了新的要求。要求企业提供的营销活动中客户的价值最大,即不仅要增大客户价值中客户的数量,更要增大客户名单中价值高的客户的数量。为了更有针对性的开展营销,企业须对那些价值比较高的客户进行更好的服务,准确的预测重要客户,正确确立目标市场是企业客户关系管理的关键。

随着信息技术的发展,商业智能系统开始出现企业信息化市场,企业的数据处理能力大大的增强,海量数据的数据仓库,先进的OLAP联机分析技术,都为数据挖掘创造了前提条件,而商务智能系统的客户的管理,实质上就是一个数据挖掘问题,数据挖掘技术是从大量的、不完全的、有噪声的、模糊的、随机的数据库中识别有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程。

文中利用数据挖掘中Naive Bayesian分类技术并结合K-means聚类算法来研究客户的重要性问题,以数据仓库中客户数据为对象,试图生成对当前数据有价值的模型,并进行适当地分析,找出预测结果与各

收稿日期:2009-08-08;**修回日期**:2009-11-10
基金项目:国家高技术研究发展计划(863)(2007AA04Z116);国家自然科学基金项目(70871033)
作者简介:赵 敏(1986-),男,安徽巢湖人,硕士研究生,研究方向为人工智能与数据挖掘;倪志伟,博士,教授,研究方向为人工智能与机器学习。

种因素之间隐藏的有价值的信息，有助于提高企业的客户关系管理水平，进一步为企业的发展战略提供可参考的依据。

1 K-means聚类与Naive Bayesian分类器

1.1 K-means聚类

K-means聚类算法属于聚类分析方法中一种基本的且应用最广的划分方法，是一种在无类标号数据中发现簇和簇中心的方法。算法接受输入量 K，然后将 N 个数据对象划分为 K 个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”来进行计算的。

算法描述如下：

输入：簇的数目 K 和包含 N 个数据对象。

输出：K 个簇，使准则函数获得最小。

Function K - means()

(1) 随机地选择 K 个数据对象，每个数据对象初始地代表一个簇的平均值或质心；

(2)(第一循环阶段) 根据簇中数据对象的平均值，将每个数据对象(重新) 赋给最近似的簇；

(3)(第二循环阶段) 更新每个簇的平均值，即计算每个簇中数据对象的平均值；

(4) 反复执行(2)、(3) 步骤，直到准则函数收敛(准则函数通常采用的是：属于 K 个聚类类别的全部数据对象与其相应的簇中心的距离平方和，并使其最小化)。

1.2 Naive Bayesian算法的分类过程

Naive Bayesian算法基于贝叶斯定理。贝叶斯定理讲述如何通过给定的训练样本集预测未知样本的类别，它的预测依据就是取后验概率。贝叶斯分类模型是一种典型的基于统计方法的分类模型。贝叶斯定理是贝叶斯理论中最重要的一个公式，是贝叶斯学习方法的理论基础，它将事件的先验概率与后验概率联系起来，利用先验信息和样本数据信息确定事件的后验概率[1]。Naive Bayesian分类算法将训练实例集分解成属性向量 A 和决策类别变量 H，假定属性向量的各分量相对于决策变量是相对独立的，也就是说各个分量独立地作用于决策变量。通过对分类算法的比较研究，Naive Bayesian分类算法可以与决策树和神经网络分类算法相媲美，表现出了高准确率和高速度。

(1)给定一个没有标号的数据样本 X，用 n 维特征向量 $x=\{x_1,x_2,\cdots,x_n\}$ 表示，分别描述 x 在 n 个属性 $\{a_1,a_2,\cdots,a_n\}$ 上的属性值。假定有 m 个类 $\{c_1,c_2,\cdots,c_m\}$，那么将样本 x 分配给 c_i 的条件就是：

$$P(c_i \mid x) > P(c_j \mid x)\ (1 \leqslant j \leqslant m, j \neq i)$$

即假定样本为类 c_i 的概率大于假定为其他类的概率。根据贝叶斯定理：

$$P(c_i \mid x) = \frac{P(x \mid c_i)P(c_i)}{P(x)}$$

其中 $P(x)$ 指的是任意一个对象符合样本 X 的概率，对于所有类来说，它是一个常数，由公式可以看出，只要使 $P(x \mid c_i)P(c_i)$ 最大即可。$P(c_i)$ 为任意一个对象为类 c_i 的概率，可以用 $P(c_i)=s_i/s$ 来计算，其中 s_i 是类 c_i 中训练样本数，s 是训练样本总数。

(2) 给定样本的类标号，假定各属性值相互条件独立，这样 $P(x/c_i)$ 可以用计算公式：

$$P(x \mid c_i) = \prod_{k=1}^{n} P(x_k \mid c_i)$$

概率 $P(x_k \mid c_i)$ 可以用训练样本估算：

如果 a_k 是离散属性，则 $P(x_k \mid c_i) = s_{ik}/s_i$，其中 s_{ik} 是属性 a_k 上值为 x_k 的类 c_i 中的训练样本数，s_i 为 c_i 中的训练样本数。

如果 a_k 是连续值属性，通常该属性服从正态分布，并把类条件概率密度函数：

$$P(x_k \mid c_i) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

其中 μ、σ 分别为属性 a_i 取值的平均差和标准差。

(3) 对未知的数据项 X 进行分类，对于每个 C，计算 $P(x \mid c_i)P(c_i)$，当且仅当 $P(x \mid c_i)P(c_i) > P(x \mid c_j)P(c_j), 1 \leqslant j \leqslant m, j \neq i$[2]。

2 采用聚类算法和Naive Bayesian分类器的客户分类模型

系统的应用模型如图1所示，首先抽取客户数据，进行预处理后基于RFM模型使用K-means聚类算法初步确定客户的价值属性，存入客户价值案例库，再对案例库中的先验数据使用Naive Bayesian分类算法验证后验概率，依据概率的大小验证K-means聚类，进行学习，并反馈入客户价值案例库。

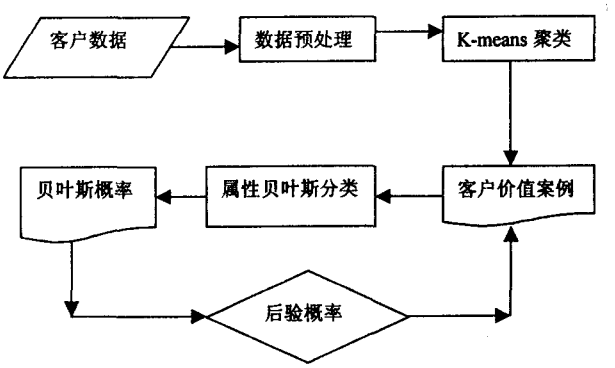

图1 客户价值分类预测学习模型

2.1 RFM模型在客户关系管理上的应用

Goodman提出了RFM模型[3],RFM模型具有三个参数:最近一次消费时间(Recency)、消费频(Frequency)、消费金额(Monetary)三个因素是衡量客户价值和客户创利能力的重要工具和手段。在RFM模式中,R(Recency)表示客户最近一次购买的时间有多远,F(Frequency)表示客户在最近一段时间内购买的次数,M(Monetary)表示客户在最近一段时间内购买的金额。一般的分析型CRM着重在对于客户贡献度的分析,RFM则强调以客户的行为来区分客户。

RFM非常适用于生产多种商品的企业,而且这些商品单价相对不高,也适合在一个企业内只有少数耐久商品,但是该商品中有一部分属于消耗品[4]。

客户价值CV可以采用下列公式来计算:

$$CV_i = \lambda_r R_i + \lambda_f F_i + \lambda_m M_i \quad (1)$$

其中$\lambda_r, \lambda_f, \lambda_m$分别为RFM中三个参数的权重系数,由于最近一次购买时间、购买总金额与消费频率对客户价值中占的比重是不同的,在这里分别设$\lambda_r = 0.2$,$\lambda_f = 0.45, \lambda_m = 0.35$,$F$对客户价值影响最大,为0.45,$M$次之,为0.35,$R$最小,为0.2,这样可以根据(1)式来计算客户价值。

2.2 利用K－means聚类算法对训练集客户数据进行聚类

通过上述的RFM指标计算出训练集客户的价值指标,然后利用K－means聚类算法来聚类是通过对数据对象集合进行分析,根据对象之间的相似度将数据对象划分成多个类,使同一类中的对象之间具有较高的相似度,不同类中的对象相异度最大,采用聚类方法可以根据客户价值的分布情况进行较为科学的划分,避免了人为划分可能带来的主观因素的影响。

从数据仓库的客户属性中选取EFM模型需要的三个属性,数据预处理后再使用K－means聚类将客户划分等级。

K－means聚类算法根据计算数据节点之间的距离来确定对象之间的相似度,在应用到客户价值聚类时,计算客户价值之间的距离,即公式(2)

$$\text{Distance}(i,j) = |CV_i - CV_j| \quad (2)$$

使用K－means聚类算法对客户关系的聚类详细的步骤如下:

(1) 选择K个客户的CV值作为聚类的中心点;

(2) 使用Distance(i,j)将客户划分到最近的类;

(3)更新每个类的平均值,计算每个类中数据对象CV的平均值,即每个类的CV值之和再除以类中客户结点数目,作为客户类的新中心点;

(4)重复2,3两步,直到各类的中心点不再发生变化[5]。

在这里,取K值为2,依据各类到中心点的距离,将训练集客户分为高价值与低价值两类,分别用H, L代表高价值客户与低价值客户,这样就可以为下一步的工作,即使用Naive Bayesian分类器预测新的客户价值。

2.3 Naive Bayesian算法对新的未知客户数据进行预测分类

分类是在顾客数据仓库中运用数据挖掘技术的一项重要任务,分类的目的是预测所给数据项的分类标号,Naive Bayesian分类器可以预测类的成员关系的概率,例如给定样本属于某一类的概率,Naive Bayesian是一种分类监督学习方法,理论上应用Naive Bayesian的前提是样本的属性独立于样本的分类属性[6]。

文中,样本是数据仓库中的海量顾客数据记录,这些样本是由顾客属性值与类别组成的多维特征向量,每个具体的样本表示方法为$\{c_i, x_1, x_2, \cdots, x_n\}$,其中$x_i (1 \leqslant i \leqslant n)$表示样本的属性值,这里用前面提到的RFM模型的三个要素来描述客户的重要性程度,即样本被表示为$\{r, f, m, c_i\}$,这里c_i表示样本所属于的类。

对客户数据进行预测分类,样本原始属性集$X = \{c_i, x_1, x_2, \cdots, x_n\}$,分类属性$C = \{c_1, c_2, \cdots, c_m\}$,每个样本可以表示为$\{C_i, 0.45R, 0.35F, 0.2M\}$。由于要把客户分为高价值客户(High value customer)与低价值客户(Low value customer),取m为2,即C_H与C_L。

对于一个未知分类号的顾客数据样本X,Naive Bayesian分类法将预测X属于具有最大后验概率(条件X下)的类,即朴素贝叶斯分类将未知的样本分配给类值独立于样本的分类属性[7]。

根据上文中由K－means聚类得出高价值客户与低价值客户的先验概率,分别用H, L代表高价值客户与低价值客户,这样就可以得出先验概率$P(C_H)$与$P(C_L)$分别为:

$$P(C_H) = \frac{C_H}{X}, P(C_L) = \frac{C_L}{X}$$

给定一个未知类标号的数据样本X,分类法将预测X属于具有最大后验概率(条件X下)的类[6],即是说,朴素贝叶斯分类将未知的样本分配给类的值独立于样本的分类属性。

$$P(c_i \mid X) = P(c_i \mid x) = \frac{P(x \mid c_i)P(c_i)}{P(x)}$$

现在,需要计算最大的$P(x \mid c_i)$即可。对样本的类,计算$P(C_H \mid X)$与$P(C_L \mid X)$的值,当且仅当$P(x \mid c_i)P(c_i) > P(x \mid c_j)P(c_j), 1 \leqslant j \leqslant m, j \neq i$。样本

分别被划分进入高价值客户与低价值客户。

由于每个客户都假定成独立的，并且结合 K－means 聚类使得在划分客户时没使用主观的判定 $P(C_H)$ 与 $P(C_L)$ 相等，Naive Bayesian 分类器在处理时具有最小的错误概率，且其学习效率很高，因此在实践中有广泛的应用价值[8]。

3 相关实验数据及结论

文中依据制造业商务智能系统数据仓库中的数据，从大量客户关系数据集中选取 3 万条客户信息数据，剔除其他属性，留下 RFM 模型的三要素，先进行数据预处理，使用 K－means 对 RFM 客户属性数据进行聚类分析，初步确定客户的价值属性，将客户划分为高价值客户与低价值客户，然后再使用 Naive Bayesian 分类器验证后验概率。

在系统的实现中，与单纯的朴素贝叶斯分类器进行比较。K 的取值可以有很多种，文中做了 K 值为 2 的情况（见表 1），即将客户只分为两个等级，即高价值客户与低价值客户，也可以根据客户关系管理的粒度粗细和实际应用分为任意个层次，比如 $K = 3$ 将客户划分为高、中、低三个层次。先使用 K－means 聚类保证了对客户的划分并不是主观上认为高价值客户与低价值客户是相等或者随意划分的，这样就排除了主观性，防止将低价值客户与高价值客户混淆，造成错误的判断。Naive Bayesian 又有一个前提条件，即它要求组成数据库的各个属性在给定类的取值中必须是互相独立的，也就是说，任何属性的取值都不依赖于其他属性。

文中采取的 RFM 模型三个要素基本相互独立，没有很强的依赖性，因而 Naive Bayesian 分类验证保证了结果的准确性和科学性，分类的准确率也比较好。

表 1 对客户进行 $K = 2$ 的 K－means 聚类（分为高价值与低价值两种）

	$K = 2$ 的 K－means 聚类
高价值客户	34.6%
低价值客户	63.4%

从以上的分析与实验得知（见表 2），结合 K－means 算法的 Naive Bayesian 分类不仅仅消除了客户划分时的主观性，而且利用了 Naive Bayesian 的高准确性，与单纯的 Naive Bayesian 分类器相比，结合 K－means 的 Naive Bayesian 分类器具有更准确的效果。

表 2 结合 K －means 的 Naive Bayesian 分类器与 Naive Bayesian 的比较

	$K = 2$ 时结合 K－means 的 Naive Bayesian 分类器	普通 Naive Bayesian 分类
准确率	78.3%	76.8%

4 结束语

文中将 K－means 聚类算法和 Naive Bayesian 技术运用于商务智能中客户价值预测中，提高了对客户价值评价的科学性、客观性和准确性，从而进一步提高了客户关系管理水平。

参考文献：

[1] 黄友平. 贝叶斯网络研究[D]. 北京：中国科学院计算技术研究所，2005：55－58.

[2] Goodman J. Leveraging the customer database to your competitive advantage[J]. Direct Marketing, 1992, 55(8): 26－27.

[3] 赵晓煜，黄小原. 基于数据挖掘的客户价值预测方法[J]. 东北大学学报，2006，12(4)：2－4.

[4] 王 珊. 数据仓库技术与联机分析处理[M]. 北京：科学出版社，1999：167－168.

[5] Kowadlo G. Improving the robustness of naive physics airflow mapping, using Bayesian reasoning on a multiple hypothesis tree[J]. Robotics and Autonomous Systems, 2009(3): 12－13.

[6] 余瑞康，施润身. 聚类思想在贝叶斯算法中的应用[J]. 计算机工程与应用，2006，42(3)：159－160.

[7] Berry M J A, Linoff G S. Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management[M]. Beijing: China Machine Press, 2006: 103－122.

[8] Han Jiawei, Kamber M. 数据挖掘：概念与技术[M]. 北京：机械工业出版社，2005：14－17.