DOI:10.3969/j.issn.1673-4785.201404050
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.tp.20150527.1021.001.html

基于细精度关联规则挖掘的电信客户流失分析

梁路,王彪,王剑辉,刘冬宁
(广东工业大学 计算机学院,广东 广州 510006)

摘 要:用决策树等常规关联规则方法分析电信客户流失问题时,存在属性相关性不够精细的问题,即无法剖析属性的内在结构、内涵及隐藏的细粒度的相关规律,同时也无法满足海量电信数据分析的需求。采用细精度关联规则挖掘解决上述问题,从逻辑学角度提出用二进制编码的方法对属性进行分解,用其构造正负训练样本集,然后进行 OCAT 关联规则挖掘,并加入启发式规则加快收敛速度,以节省时间和内存开销。实验结果表明,基于这种方法产生的关联规则提高了细精度,同时易于实施并行计算和提高效率,能更好地满足当前电信应用需求。

关键词:电信客户流失;细精度;关联规则;逻辑方法;OCAT;启发式规则
中图分类号:TP182 **文献标志码**:A **文章编号**:1673-4785(2015)03-0407-07

中文引用格式:梁路,王彪,王剑辉,等. 基于细精度关联规则挖掘的电信客户流失分析[J]. 智能系统学报, 2015, 10(3): 407-413.
英文引用格式:**LIANG Lu, WANG Biao, WANG Jianhui, et al. Analysis of telecom customer churn based on fine-grained association rule mining[J]. CAAI Transactions on Intelligent Systems**, 2015, 10(3): 407-413.

Analysis of telecom customer churn based on fine-grained association rule mining

LIANG Lu, WANG Biao, WANG Jianhui, LIU Dongning
(Faculty of Computer Science, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: When using traditional association rule mining such as decision tree to analyze the problem of telecom customer churn, we always meet the problem that the dependency of attributes are not enough fine, which means traditional methods not only cannot analyze the internal structure and hidden fine-grained related rules of attributes, but also cannot satisfy the needs of analyzing massive telecom data. In this paper, we solve the above problems by using fine-grained association rule mining. We firstly design a binary coding method from logic viewpoint to break attributes to segments, and then build the positive and negative training sample sets based on segments. In experiment we adopt the one clause at a time (OCAT) algorithm on association rule mining for speeding up the convergence speed and saving the overhead of time and memory. Finally, the experimental result shows that this method improves the fine-grained of the association rule, which can be easily used in parallel computing to raise efficiency, and satisfy the requirements of current telecom application.

Keywords: telecom customer churn; fine grain; association rules; logic method; one clause at a time (OCAT); heuristic rules

当前各电信企业市场竞争越演越烈,为了提高客户忠诚度,迫切要求企业借助于对日益庞大的历史数据进行分析,制定更好的技术方案和营销策略。然而影响客户忠诚度的因素非常复杂,营销人员不通晓技术,技术人员又不精于营销,且数据挖掘是目前最有效的数据分析手段之一,用于发现大量数据所隐含的各种规律[1],因此选择一种合适的数据挖掘方法极为重要。目前常用的方法是关联规则挖

收稿日期:2014-04-27. **网络出版日期**:2015-05-27.
基金项目:国家"863"计划重大项目(2013AA01A212);国家自然科学基金资助项目(61272067, 61104156);广东省自然科学基金资助项目(9451009001002777).
通信作者:王彪. E-mail: wangbiao_gdut@163.com.

掘[2],因为其能够比较直观地得出各因素之间的关系,而且操作过程简单,结果的可解释性强[3]。但是,现有的关联规则挖掘方法均无法进一步发现隐藏在属性内部的相关规律,并且在面向海量数据挖掘时效率很低。

目前用于电信客户流失预测的方法主要可分为 3 类[4]:第 1 类方法以传统的统计学理论为基础,主要包括聚类、贝叶斯分类器、决策树和逻辑回归等。如 Kim 等[5]曾采用逻辑回归方法对韩国部分移动客户进行了流失预测分析,探讨了韩国移动电信市场相关因素在客户流失和忠诚度之间的关系,为保持客户的忠诚度提供了帮助。第 2 类方法以人工智能理论为基础,主要包括人工神经网络和进化学习等。如 Mozer 等[6]曾采用人工神经网络方法结合数据抽样等方法建立了客户流失预测模型,并为某电信公司进行了客户流失预测,通过与决策树等方法对比,发现采用该方法产生的关联规则预测效果更好,准确率更高。第 3 类方法以统计学习理论为基础,其典型代表为支持向量机方法。如邝涛等[7]采用基于代价敏感学习的支持向量机模型对某电信公司的客户数据进行挖掘,并通过与神经网络等方法对比,发现该方法能获得较高的预测精度和覆盖率,并能在某种程度上解决了数据集非平衡性等问题。

尽管以上 3 类方法被大量使用,但它们都忽略了属性内在结构之间的细粒度的相关规律,即存在关联规则不够精细的问题;并且在用于海量数据分析时计算量大、效率低,难以及时反映客户的流失倾向,因此不能完全满足当前电信应用的需求。因此,本文的思路是把每个独立的属性“打碎”,即分解得到细粒度的“属性片段”,以提高关联规则的 fine 精度[8],再基于属性片段采用合适的关联规则挖掘算法,最后得到的关联规则要易于对海量数据实施并行计算提高效率,从而可以更好地及时定位影响客户流失的关键因素,或发现一些隐藏的关键规律,使其能在电信客户流失预测中具有更大的应用空间。

1 电信数据挖掘分析与改进

为了得到属性内在相关规律及方便实现并行运算,采用了基于逻辑的细精度关联规则挖掘方法。首先提出了与领域相关的属性分解方法,即对属性值域进行合适的分类以得到“属性片段”,再对每个分类进行二进制编码。而这样变化后的数据并不适合采用决策树和聚类等传统方法进行处理,于是我们结合了 E. Triantaphyllou 提出的基于逻辑的 OCAT 方法[9]35-45进行关联规则挖掘。这种方法的核心是通过某种基于逻辑的方法寻找一系列子式,再由这些子式合取得到关联规则表达式,该表达式由二进制化后的“属性片段”构成,不仅能直观地展现属性的内部结构和内涵,而且这种由合取范式表示的关联规则特别适合采用并行计算进行海量数据分析,从而提高运算效率。图 1 展示了该方法的具体步骤:数据预处理(数据清洗、数据二进制编码和正负样本构造)、关联规则的挖掘、结果检验以及评估反馈。

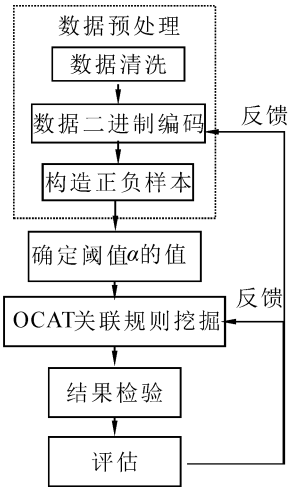


图 1 细精度关联规则挖掘流程

Fig. 1 Process of fine-grained association rule mining

1.1 数据预处理

文中数据集为某电信运营商某地区数据仓库中的客户原始历史数据,总共有 176 921 条记录,其中正常客户记录为 156 885 条,流失客户记录 20 036 条。

1)数据清洗与属性变换。

首先删除有效值较少的属性或记录,以及非法的记录。其次,在电信客户流失的应用背景下,需要研究能够反映客户消费承受能力、消费习惯和消费行为的属性相关性,同时用文献[10]的经验做支撑,为了能够准确地覆盖影响客户流失的因素,构建了月消费比率 r_i(上一个月与本月的消费比值)、年度本地通话费用(year_local_fee)、未消费月份数(non_fee)3 个属性。在实验中选用了 r_1(2008 年 12 月的消费额/2009 年 1 月的消费额)、r_2(2009 年 1 月的消费额/2009 年 2 月的消费额)f_{0902}(2009 年 2 月的消费额)、year_local_fee 等 5 个属性作为特征属性。

2)数据二进制编码。

考虑到属性值域分类的语义差别,及其与客户流失问题的相关性差别,采用了与领域相关的二进制编码方法将属性“掰碎”,即根据记录的取值相关性进行分类。经过分析,发现实验用到的电信用户数据集具有典型的密度特性,因此采用了基于密度

的可以快速发现任意形状类的 DBSCAN 分类算法[11]完成从“属性”到“属性片段”的分解。另外,此数据集不同属性值域区间的差异较大,这可能会影响分类的效果,因此还需要对这些属性进行常用的归一化处理。具体的步骤是:首先使用 mapminmax(x,min,max)函数将值域归一到区间[0,1],其次以每个属性数据结合函数 ones(n,1)产生的数据作为输入 x(其中 n 为样本的个数),通过经验以及多次实验确定较优的参数。结果集中可能会存在孤立点,较优的参数能够减少孤立点的个数,并且产生较为合理的分类。在实际应用中,孤立点的数量往往远小于正常数据[12],因此其影响比较小,故本文就近将孤立点分配至临近的区间(相应的孤立点处理方法作为后续工作)。经过 DBSCAN 运算后,便得到了属性分类,即“属性片段”,这些分类结果直接反映了其属性值自然特有的内在规律,由事物本身决定,不受人为控制,因此对后续进一步分析隐藏在它们之间的相关性有重要的指导作用。最后,再根据分类的个数,使用若干二进制位对属性片段进行二进制编码。以属性 year_local_fee 为例,其被划分为 7 个区间,包含 65 个孤立点,将这些孤立点按照就近原则并入 7 个区间内,故以 3 位二进制位来表示,如表 1 所示。实验中各属性的分类数如表 2 所示。

表 1　属性 year_local_fee 的二进制编码对应关系

Table 1　Classification of year_local_fee and its binary encoding

区间	1	2	3	4	5	6	7
编码	000	001	010	011	100	101	110

表 2　所有属性的分类数

Table 2　Classification quantity of all attributes

属性	分类数
non_fee	3
year_local_fee	7
r_1	7
r_2	5
f^e_{0902}	8

3)正负样本的构造。

首先,根据数据集中 churn 标志位(标志客户流失信息的属性)的值,将整个数据集的用户划分为已经流失的客户(churn 值为 0)和未流失的客户(churn 值为 1)2 类。其次,为了得到关联规则挖掘中需要的训练集和检验集,实验分别随机抽取未流失客户和已流失客户中四分之三的数据[13]来构造训练集数据的正负样本(用 E^+ 和 E^- 来分别表示正样本和负样本),剩余的数据用作检验集。最后,通过前后 10 次进行随机抽取得到正负样本,并分别计算结果,进一步验证数据挖掘结果的稳定性、可信性和鲁棒性。在构造 E^+ 和 E^- 时,采用了如表 3 所示的属性排列顺序及其编码对应关系。

表 3　属性与二进制编码对应关系

Table 3　Binary encoding of all attributes and its representation

属性	二进制编码
non_fee	A_1
	A_2
year_local_fee	A_3
	A_4
	A_5
r_1	A_6
	A_7
	A_8
r_2	A_9
	A_{10}
	A_{11}
f^e_{0902}	A_{12}
	A_{13}
	A_{14}

1.2　OCAT 关联规则挖掘

经过数据预处理得到的二进制数据不能采用决策树和聚类等传统方法处理,同时为了让关联规则易于实施并行计算,实验中采用了 OCAT 方法,该方法每次产生一条最优子式(接受全部正样本,拒绝尽可能多的负样本),最终将子式通过合取操作得到关联规则。但是,OCAT 产生单个子式的过程中剪枝的数量较少,导致收敛速度缓慢,并且需要存储大量的叶子结点的限界,导致运算时耗费了极大的时间和空间。因此,实验中最终引入了启发式规则[9]73-80进行快速剪枝,将时间复杂度从指数级别降到多项式级别,这将非常适合数据量庞大的电信数据挖掘。

启发式方法中根据 POS(a_k)/NEG(a_k)的比值逆序排列属性(片段)a_k(a_k 为 A_i 或 $\neg A_i$)形成有序集合 A,挑选 A 中前 α(%)(阈值)的元素生成待选集合 L。关联规则的构造过程即不断地从待选集合 L 中随机选出属性(片段)a_k 加入子式的过程,该过程持续到关联规则完成或集合 L 为空。若 α 值过小,则关联规则不完整,导致预测结果准确度下降;反之,若 α 值过大,则对客户流失影响很小的属性(片段)可能会包含在关联规则中,从而导致关联规则

的子句过长、数量过多,以及属性间的相关性不够强。由于二进制编码方法和关联规则挖掘过程的随机性都会影响所生成的关联规则的准确性,故需要在实验过程中进行多级反馈和修正。基于启发式的 OCAT 算法流程如下。

输入:正负样本 E^+和 E^-。 //E^+、E^-分别是未流失和已流失客户数据集。

输出:程序执行多次所得最优的关联规则 C,其规范形式为式(1)所示,其中析取子式由属性片段析取得到,并能保证子式的数量 n 最少,且子式中属性片段的数量也是最少。

$$C = (\text{析取子式 } 1) \wedge \text{析取子式 } 2 \wedge \ldots \wedge \text{析取子式 } n \wedge \cdots \quad (1)$$

初始化 E^+、E^-和 $C = \varnothing$

Do While ($E^- \neq \varnothing$)

重置集合 A,$C_i = \varnothing$; //A 为二进制编码属性片段 a_k的集合,a_k为 A_i或$\neg A_i$

Do While ($E^+ \neq \varnothing$)

1)根据 POS(a_k)/NEG(a_k)的比值逆序排列 A 中的元素 a_k(如果 NEG(a_k)为零,则将 POS(a_k)作为它的值);

2)选择有序集 A 中前α(%)的元素生成待选集合 L;

3)从 L 中随机选择属性 a_k加入到 C_i中;

4)$E^+ \leftarrow E^+ - E^+(a_k)$; //$E^+(a_k)$为包含 a_k时 C_i所能接受的 E^+的中的元素集合

5)$A \leftarrow A - a_k$; //将 a_k从集合 A 中剔除

6)对所有 $a_k \in A$ 重新计算 POS(a_k)的值;

Repeat

7)$C \leftarrow C_i \wedge C$; //将子式 C_i合取到关联规则 C 中

8)$E^- \leftarrow E^- - E^-(C)$; // $E-(C)$为 C 目前所能拒绝的 E^-中的元素集合

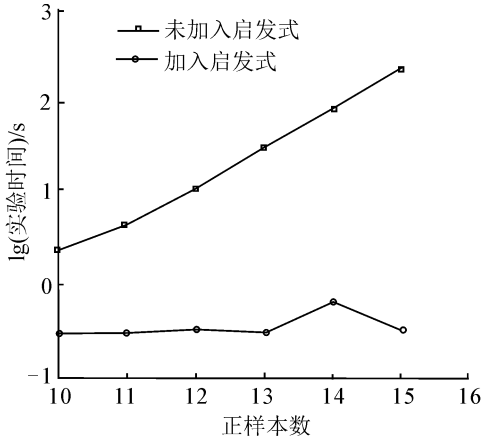
9)重置 E^+;

Repeat

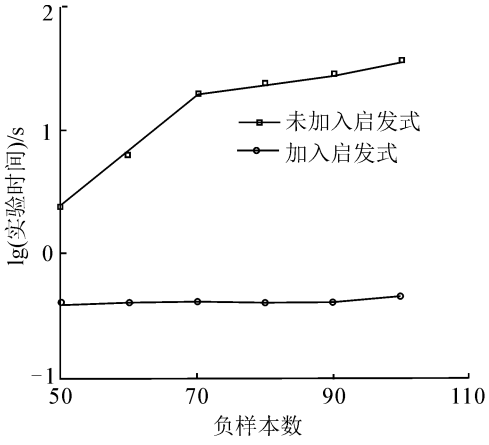
2 细精度关联规则挖掘

为了验证上述方法在预测客户流失过程中的可行性及有效性,我们设计并实现了各个过程。实验环境为 1) CPU/内存/硬盘: AMD Athlon(tm) Ⅱ X2215/DDR2 4 GB/320 GB 7 200 转/min;2)平台/环境/语言 Windows 8.1 64 bit 操作系统、Microsoft Visual Studio 2013/C、C#。

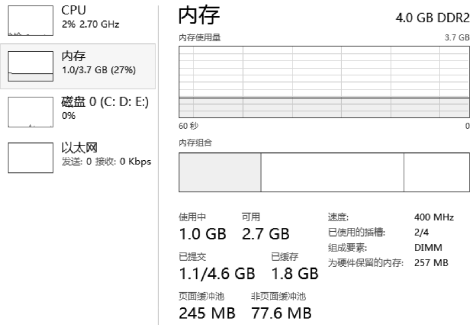
为了验证加入启发式规则后算法的收敛效果,在相同正负样本和相同环境下进行了时间与空间消耗的对比实验。通过多次实验,选取的负样本基数为 50,这种大小的样本空间可以让 OCAT 方法的耗时不会太大,又能明显比较出 2 种方法在相同数据集上的时间耗费差异。在改变正样本的基数时,分别使用上述 2 种方法的运算时间对比如图 2(a)所示。同理,将正样本基数固定为 10,负样本基数从 50 开始,每次增加 10 条负样本记录,一直到 100 条,此时得到的运算时间对比如图 2(b)所示。此外,在 300 条正样本与 100 条负样本情况下,对二者的内存占用情况进行对比,其中图 2(c)为未运行程序时的内存占用,图 2(d)与图 2(e)分别为使用 OCAT 方法与加入启发式方法后的内存占用情况。



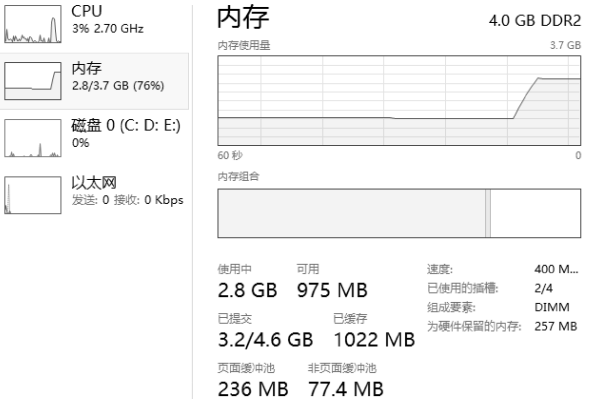
(a)正样本数-时间关系



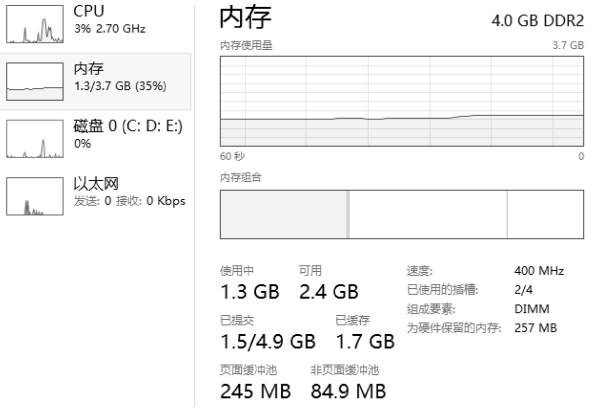
(b)负样本数-时间的关系



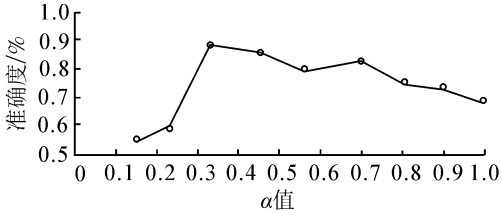
(c)电脑空闲时的内存



(d)未加入启发式时内存



(e)加入启发式后内存



(f)α 值-准确度的关系

图 2　实验结果

Fig. 2　Experiment results

从图 2(a)~(e)可以看出,加入启发式规则之后的方法比 OCAT 方法所消耗的时间与空间都大大减少,图 2(f)描述了不同的阈值 α 值对预测准确度的影响,体现了其重要性。最后,10 次实验所得的关联规则之一如式(2)所示。

$$C = (\neg A_1 \vee A_3 \vee A_5 \vee A_6 \vee \neg A_9 \vee A_{12}) \wedge \cdots \wedge (A_4 \vee \neg A_5 \vee A_8 \vee A_9 \vee \neg A_{11} \vee \neg A_{13} \vee A_{14}) \wedge (A_2 \vee \neg A_3 \vee \neg A_6 \vee A_7 \vee \neg A_{10} \vee A_{12} \vee \neg A_{14}) \wedge (A_1 \vee \neg A_5 \vee \neg A_{10} \vee A_{13}) \quad (2)$$

该表达式为关联规则的布尔表达式,它由 4 条析取范式的子式经过合取操作得到,其中每条子式均能接受所有的正样本而拒绝若干负样本,所有子式合取而成的关联规则能接受所有正样本并且拒绝所有负样本。这一结果具有如下特点:

1)反映了一些新的规律。以子式($\neg A_1 \vee A_3 \vee A_5 \vee A_6 \vee \neg A_9 \vee A_{12}$)为例,对照表 1 分析得出属性 non_fee 与其他属性 Year_local_fee、r_1、r_2、f_{0902}^e的“内在片段”的相关性,即“属性+片段->片段”或“片段+片段->属性”,而传统的关联规则只能发现属性与属性的相关性,即“属性+属性->属性”。

2)运算效率高。以子式($\neg A_1 \vee A_3 \vee A_5 \vee A_6 \vee \neg A_9 \vee A_{12}$)为例,其仅有 6 个原子项,因此运行较快。而由于上述关联规则为合取范式,以“∧”为连接词构成,因此用这种形式的规则检验海量数据时可采用高度并行计算的方法,进一步了减少时间开销。

3)剖析了属性内部结构与内涵。以子式($\neg A_1 \vee A_3 \vee A_5 \vee A_6 \vee \neg A_9 \vee A_{12}$)为例,其等价于($A_1 \wedge \neg A_3 \wedge \neg A_5 \wedge \neg A_6 \wedge A_9$) $\Rightarrow A_{12}$,体现的已非属性之间的关联,而是属性片段与片段的相关性,即片段 A_{12}的取值取决于片段 A_1、A_3、A_5、A_6和 A_9。这一结果有利于公司决策者制定更为精准的市场策略。

4)提高了 fine 精度。10 次实验所得到的关联规则分别对应 10 条主合取范式,这些主合取范式可以使用矩阵形式表示,式(2)的主合取范式如表 4 所示。其余 9 条主合取范式与表 4 所示的主合取范式的相似度在 86.4%~90.2%,体现了结果的稳定性、可信性和鲁棒性,说明了提升 fine 精度的合理性。

表 4　主合取范式形式

Table 4　Principal conjunctive normal form

编码	SubF1	SubF2	SubF3	SubF4
A_1	0	△	△	1
A_2	△	△	1	△
A_3	1	△	0	△
A_4	△	1	△	△
A_5	1	0	△	0
A_6	1	△	0	△
A_7	△	△	1	△
A_8	△	1	△	△
A_9	0	1	△	△
A_{10}	△	△	0	0
A_{11}	△	0	△	△
A_{12}	1	△	1	△
A_{13}	△	0	△	1
A_{14}	△	1	0	△

注:符号“△”表示取 0 和 1 均可,SubF1 有 256 条,SubF2 有 128 条,SubF3 有 128 条,SubF4 有 1 024 条

5)得到了更直观、清晰的语义解释。以子式($A_1 \vee \neg A_5 \vee \neg A_{10} \vee A_{13}$)为例,若某一客户数据使该

子式结果为 0($\neg A_1$、$\neg A_5$、$\neg A_{10}$、A_{13}取值均为 0),则可以预测该用户为流失客户。根据 A_1、$\neg A_5$、$\neg A_{10}$、A_{13}的取值映射到表 3,可以得出在 non_fee 的 3 区间、year_local_fee 的 2,4,6 区间、r_2的 3,4 区间与 f^e_{0902}的 1,3,5,7 区间共同影响客户流失,根据区间与属性值对应的关系可知,客户数据符合表 5 取值的均为流失客户。

表 5 流失客户的数据特征

Table 5 Data characteristics of losing customers

属性	取值范围
non_fee	6~12
year_local_fee	30~50
	70~88
	92~133
r_1	0.320~0.416
r_2	0~20
	25~29
	34~50
	48~68
f^e_{0902}	6~12

3 结束语

针对决策树等常规关联规则方法在电信客户流失预测中遇到的属性相关性不够精细,处理大规模数据运算效率低的问题,本文采用了基于逻辑的细精度关联规则方法。该方法从逻辑学角度,通过与领域相关的二进制化技术对属性进行分解,并用得到的二进制数据构造训练集的正负样本,再使用 OCAT 方法对正负样本进行挖掘得出关联规则。然而,实验过程中耗费了极大的时间与空间,这表明直接用该方法进行海量电信数据的挖掘是不理想的。因此引入了启发式规则对其进行改进,将时间复杂度从指数级别降低到多项式级别。最后通过实验结果分析,验证了该方法能进一步体现属性的内在结构、内涵及隐藏的细粒度的相关规律,提高了关联规则的 fine 精度,并且这种由合取范式表示的关联规则特别适合实施并行计算,有利于大规模电信数据的处理,因此该方法是满足目前电信行业需求的一种较理想的数据挖掘方法。

尽管上述方法取得了不错的效果,但是对于不同数据集在具体应用时还存在一些困难,如如何更好地结合领域知识和数学方法对属性进行分解及二进制编码进而构造正负样本,如何寻找更好的启发式规则提高运算性能等。另外,当数据样本比较小的时候,得到的关联规则的准确率不够高,但在数据样本足够大的情况下,关联规则预测准确率会比较理想。在今后的工作中,将努力完善本方法的各个环节,同时找到适用本方法的数据集特征,以应用到更合适的实际问题中。

参考文献:

[1]朱扬勇, 熊赟. DNA 序列数据挖掘技术[J]. 软件学报, 2007, 18(11): 2766-2781.
ZHU Yangyong, XIONG Yun. DNA sequence data mining technique[J]. Journal of Software, 2007, 18(11): 2766-2781.

[2]贺炜, 潘泉, 陈玉春, 等. 关联规则挖掘与因果关系发现的比较研究[J]. 模式识别与人工智能, 2005, 18(3): 328-333.
HE Wei, PAN Quan, CHEN Yuchun, et al. A Comparison between association rule data mining and causal discovery [J]. Pattern Recognition and Artificial Intelligence, 2005, 18(3): 328-333.

[3]毛宇星, 陈彤兵, 施伯乐. 一种高效的多层和概化关联规则挖掘方法[J]. 软件学报, 2011, 22(12): 2965-2980.
MAO Yuxing, CHEN Tongbing, SHI Bole. Efficient method for mining multiple-level and generalized association rules [J]. Journal of Software, 2011, 22(12): 2965-2980.

[4]夏国恩. 客户流失预测的现状与发展研究[J]. 计算机应用研究, 2010, 27(2): 413-416.
XIA Guoen. Research on current situation and development of customer churn prediction[J]. Application Research of Computers, 2010, 27(2): 413-416.

[5]KIM H S, YOON C H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market [J]. Telecommunications Policy, 2004, 28(9/10): 751-765.

[6]MOZER M C, WOLNIEWICZ R, GRIMES D B, et al. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry[J]. IEEE Transactions on Neural Networks, 2000, 11(3): 690-696.

[7]邝涛, 张倩. 改进支持向量机在电信客户流失预测的应用[J]. 计算机仿真, 2011, 28(7): 329-332.
KUANG Tao, ZHANG Qian. Application of telecom customer churn prediction based on improved support vector machine[J]. Computer Simulation, 2011, 28(7): 329-332.

[8]FOX C, LAPPIN S. Foundations of intensional semantics [M]. New York: Wiley-Blackwell, 2008: 78-82.

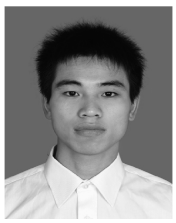
[9]TRIANTAPHYLLOU E. Data mining and knowledge discovery via logic-based methods: theory, algorithms, and appli-

cations[M]. New York: Springer, 2010.
[10]蒋盛益, 李霞, 郑琪. 数据挖掘原理与实践[M]. 北京: 电子工业出版社, 2013: 211-212.
[11]王鑫, 王洪国, 王珺, 等. 数据挖掘中聚类方法比较研究[J]. 计算机技术与发展, 2006, 16(10): 20-22.
WANG Xin, WANG Hongguo, WANG Jun, et al. Comparison of clustering methods in data mining[J]. Computer Technology and Development, 2006, 16(10): 20-22.
[12]张净, 孙志挥, 杨明, 等. 基于网格和密度的海量数据增量式离群点挖掘算法[J]. 计算机研究与发展, 2011, 48(5): 823-830.
ZHANG Jing, SUN Zhihui, YANG Ming, et al. Fast incremental outlier mining algorithm based on grid and capacity [J]. Journal of Computer Research and Development, 2011, 48(5): 823-830.
[13]胡文瑜, 孙志挥, 吴英杰. 数据挖掘取样方法研究[J]. 计算机研究与发展, 2011, 48(1): 45-54.
HU Wenyu, SUN Zhihui, WU Yingjie. Study of sampling methods on data mining and stream mining[J]. Journal of Computer Research and Development, 2011, 48(1): 45-54.

作者简介:

梁路,女,1980 年生,副教授、博士,中国计算机学会协同计算专业委员会委员。主要研究方向为协同计算、云计算和数据挖掘。主持和参与国家级、省级自然科学基金及科技计划项目,以及校企合作产学研项目多项。2011 年获广东省科学技术二等奖。发表学术论文 30 余篇。

王彪,男,1989 年生,硕士研究生,主要研究方向为数据挖掘及协同计算。



王剑辉,男,1990 年生,硕士研究生,主要研究方向为数据挖掘。