《智能信息处理》课程作业

基于形式概念的电信企业客户流失分析

李锦峰

作业	分数[20]
得分	

2020 年 12 月 06 日

基于形式概念的电信企业客户流失分析

李锦峰

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘 要 在现代市场中，已经完全转变为买家市场，面对日益激烈的市场竞争，企业开发一个新客户的成本比留住一个老客户要大得多。虽然企业经营者越来越懂得客户对企业得重要性，但是由于缺乏有效的数据分析，客户流失现象依然经常发生。本文将概念格应用于电信企业客户流失数据的分析，对电信企业客户流失的原因进行详细的分析，将客户流失的数据用概念格表示，根据客户的基本信息、电信企业提供的服务信息、客户的使用信息对客户流失数据中的主要属性集信息进行提取，将提取出的对象集和属性集的对应关系表示成形式背景，对其形式背景进行标准化，对标准化的形式背景完成客户流失信息概念格的构造，最后根据概念格的特性和概念格生成的关联规则，进行客户流失数据中的数据挖掘，通过对客户流失数据的数据挖掘给电信企业提供辅助决策的支持，预防客户的流失。

关键词 形式概念分析 概念格 数据挖掘 预防客户流失

Data Analysis of Customer Churn in Telecom Enterprises Based on Formal Concept

LiJinFeng

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract In the modern market, it has been completely transformed into a buyer's market. In the face of increasingly fierce market competition, the cost of a company to develop a new customer is much greater than to retain an old customer. Although business operators are more and more aware of the importance of customers to the company, due to the lack of effective data analysis, customer churn still occurs frequently. This paper applies the concept grid to the analysis of customer churn data in telecommunications companies, and conducts a detailed analysis of the causes of customer churn in telecommunications companies. Use the information to extract the main attribute set information in the customer churn data, express the corresponding relationship between the extracted object set and the attribute set as a formal background, standardize its formal background, and complete the concept of customer churn information on the standardized formal background According to the characteristics of the concept lattice and the association rules generated by the concept lattice, data mining is carried out in the customer churn data. Through the data mining of the customer churn data, it provides support for decision-making assistance to telecom companies to prevent customer churn.

Key words Formal concept analysis; Concept lattice; Data mining; Prevent customer loss;

1 引言

形式概念分析[1]是由德国 wille 教授于 1982 提出并逐步完善成一门有效的数据分析与知识发现技术。该技术用于数据分析的核心工具是概念格，概念格是知识的一种表现模型，描述了对象和特征之间的联系，表明了概念之间的泛化和例化关系，体现了概念的分类和层次关系，因此概念格作为一种有效的数据分析工具，非常适合用来发现规则型知识。关联规则是从大量的数据中挖掘有价值的知识，这些知识体现了数据项之间的相互联系[3]。

2 基本定义

在形式概念分析中，形式背景（formal context）通常被定义为一个三元组 K=（U,D,R），其中 U 是对象（实体）集合，D 是描述符（属性）集合，R 是 U 与 D 之间的一个二元关系。即 R⊆U×D.xRd 被读作“对象 x 具有特征 d”，在形式背景 K 中，在 U 的幂集之间可以定义两个映射 f 和 g 如下[5]：

$$\forall O_1 \subseteq U: f(O_1) = \{d \in D \mid \forall x \in O_1 \ (xRd)\};$$

$$\forall D_1 \subseteq D: f(D_1) = \{x \in U \mid \forall d \in D_1 \ (xRd)\};$$

它们被称为 U 的幂集和 D 的幂集之间的 Galois 联系，来自 P(U)×P(D)的二元组（O1,D1）如果满足两个条件：$O1 = g(D1)$及 $D1 = f(O1)$，则它被称为是形式背景 K 的一个形式概念 ,其中 D1 和 O1 分别被称为概念(O1 , D1)的内涵和外延. K 的所有形式概念的集合被标记为 CS(K)， CS (K) 上最重要的结构是由亚概念 -超概念关系 (又称为泛化 -例化关系 ,或前驱-后继关系)产生的 ,其定义如下:如果 O1⊆O2,则形式概念（O1,D1）是形式概念（O2,D2）的亚概念，记为（O1,D1）≤（O2,D2），通过这个关系，我们得到一个有序集 CS(K)=（CS(K),≤）,这是一个完全格，被称为形式背景 K 的概念格，概念格中的每个结点都是一个形式概念，对于概念格中两个不同的结点 C1=（O1,D1）和 C2=(O2,D2)，如果 C1 是 C2 的亚概念且不存在其他的结点 C3 满足 C1≤C2≤C3，则 C1 称为是 C2 的子结点（直接后继），而 C2 是 C1 的父结点（直接前驱）[5]。

从上述定义可以看出，每个结点（概念）C1=（O1,D1）的内涵都是最大化的：对于任意 D2⊃D1, 都有 g(D2)⊂g(D1)=O1,即对内涵的任意扩充都会使相应的外延缩小，由此，我们可以定义出概念结点内涵的最小化—内涵缩减[5]。

定义 1.对于概念格的结点（概念）C1=(O1,D1), 特征子集 D2 被称为是（O1,D1）的内涵缩减当且仅当：

$$(1)\ g(D2) = g(D1) = O1;$$

$$(2)\ \forall D3 \subset D2, g(D3) \supset g(D2) = O1;$$

其中条件（1）称为内涵缩减的外延不变性，即结点（O1,D1）的内涵 D1 和它的内涵缩减 D2 具有相同的外延，条件（2）称为内涵缩减的最小性，即从中任意去除一个属性将会导致外延的增加，C1 的所有内涵缩减的族集被标记为 IN T-RED（C1）[5]。

定理 1.对于概念格中的结点（概念）C=（O1,D1）以及特征子集 D2,有 g(D2)=g(D1)=O1,当且仅当 D2⊆D1 且对于 C 的任意父结点 C2=(O3,D3)有 D2∩（D1-D3）≠∅。[5]

定义 2.对于概念格 L（K）中，K=（U,M,I），由形式概念 C1,C2 构成的先辈晚辈节点对（C1,C2），假定形式概念 C1=（O1∪O,A）,C2=(O,A∪B),也即，在既有属性 A 的|O1∪O|对象中，有|O|个对象也具有属性 A[7]。那么，就可以得到

——关联规则 A⇒B

——关联规则 A⇒B 的支持度：

$$sup\ p(A \Rightarrow B) = \frac{|O|}{|U|} = \frac{|Extent(C2)|}{|U|}$$

——关联规则 A⇒B 的可信度：

$$conf(A \Rightarrow B) = \frac{|O|}{|O1 \cup O|} = \frac{|Extent(C2)|}{|Extent(C1)|}$$

这样在概念格 L(K)上提取关联规则时我们就可以只关心关联规则 A⇒B 中支持度和可行度大于阈值的关联规则即可。

3 基于概念格的电信客户流失数据表示

概念格这种知识模型作为知识组织和知识挖掘的有效工具，已在学术界和产业界得到了广泛的关注。同时，概念格的构建工具也被许多组织和机构利用和研究[3]。目前，常用的概念格构造器包括：Lattice Miner，ConExp 等多种，本文所用的概念格构造器是 ConExp。

利用 ConExp 构造概念格德尔步骤主要有以下各个步骤：

（1）在电信客户流失数据中提取主要的对象和属性；

（2）确定概念格的对象集和属性集；

（3）确定对象和属性之间对应的关系；

（4）生成电信客户流失数据的概念格；

3.1 电信企业客户流失数据主要信息提取

电信客户流失数据主要通过 Kaggle 的电信客户流失预测比赛中取得，该数据集中总共包含了 7040 名顾客的信息，这些信息包括了客户的基本信息，客户的使用信息，以及电信公司提供的服务信息，总共 21 个属性。本文从 7040 名顾客中随机抽取了的 20 名顾客， 并从 21 个属性中抽取了 12 个属性作为概念格的属性集。

这 12 个属性包括：客户是否是老年人，客户是否有合伙人，客户所使用服务是否有附属客户，客户使用公司服务的月数，客户是否办有电话服务，客户是否办理了多条电话服务渠道，客户的网络服务提供线路，客户是否使用网络安全服务，客户的合约方式，客户的每月支出情况，客户从使用至今的总支出情况，客户是否流失。考虑到客户使用公司服务月数以及客户每月支出情况两个属性为多值属性，为了方便之后的形式背景的标准化将两属性的取值按照其对应的平均数，分为低和高（小于平均值为低，大于等于平均值为高）。具体的属性集如图 3.1 所示。

属性集			
客户是否老年人	是(a1)	客户的网络服务提供线路	数字用户线路(g1)
	否(a2)		光纤线路(g2)
客户是否有合伙人	是(b1)		未办理网络服务(g3)
	否(b2)	是客户是否使用网络	是(h1)
客户所使用服务是否有附属用户	是(c1)		否(h2)
	否(c2)	用户的合约方式	每月签约(i1)
客户使用公司服务的月数	0-72 之间(d1,d2)		一年(i2)
客户是否办有电话服务	是(e1)		两年(i3)
	否(e2)	客户的每月支出情况	金额(j1,j2)
客户是否办理了多条电话服务渠道	是(f1)	客户是否流失	是(k1)
	否(f2)		否(k2)

图 3.1 电信企业客户流失数据属性集

3.2 电信企业客户流失数据形式背景标准化

由上节确定的对象集和属性集给出对应值信息，数字 1 表示对象具有该属性。数字 0 表示对象不具有该属性。为了简化属性集，我们将属性集中的属性分布用小写字母来表示。如下表 3-1 客户流失数据的形式背景。(受限于篇幅只列出少部分属性以及对象)

表 3-1 电信企业客户流失数据的形式背景

	a1	a2	b1	b2	…	k1	k2
U1	0	1	1	0	…	1	0
U2	0	1	1	0	…	0	1
U3	0	1	0	1	…	0	1
U4	0	1	1	0	…	1	0
U5	0	1	1	0	…	0	1
…	…	…	…	…	…	…	…
U20	0	1	1	0	…	0	1

接下来将电信企业客户流失数据的形式背景标准化，标准化后的形式背景如下图 3.2 所示。

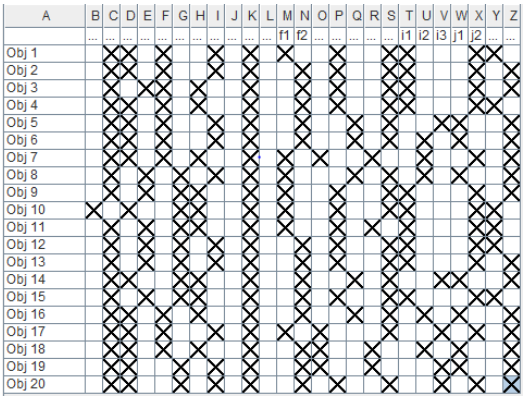

图 3.2 电信企业客户流失数据的形式背景标准化

3.3 电信企业客户流失数据概念格构造

利用概念格构造器将标准化的形式背景构造客户流失数据的概念格。属性集 24 个属性，对象集 20 个对象生成概念格，这是一个简化的概念格，一个小球代表一个概念，上半部分是蓝颜色的小球代表这个概念包含一个属性，下半部分是黑颜色的小球代表这个概念包含一个对象。如下图 3.3 所示

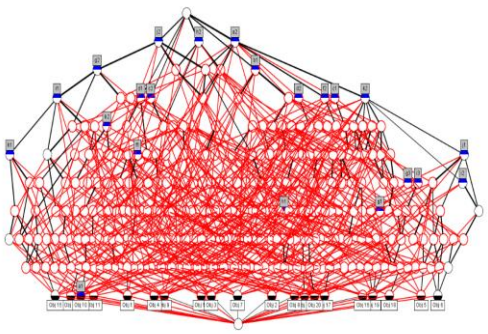

图 3.3 基于电信企业客户流失数据生成的概念格

4 基于电信企业客户流失数据概念格的数据挖掘

概念的内涵，与事务数据库中的项目集，非常类似，而且有更严格的限制，因此可在概念格上提取关联规则，并且比直接在事务数据库上提取，有更多的优势[7]。为了在概念格中发现更有意义的关联规则，设置了最小支持度和最小置信度这两个阈值，从概念格中挖出的关联规则只有满足最小支持度和最小置信度这两个阈值才具有现实意义。

根据不同的最小置信度和最小支持度的设定，我们所得到的关联规则不同，而且关联规则数量也不一样。本文设定最小支持度为 10%，最小置信度为 50%，通过 ConExp 概念格构造器提取基本的关联规则。符号“⇒”表示关联的意思。具体的说明如表 4-1 所示。

表 4-1 电信企业客户流失数据概念格中的部分关联规则

	基本关联规则	最小置信度
1	c1⇒a2	100%
2	d2⇒a2	100%
3	f2⇒a2	100%
4	k2⇒a2	100%
5	h1,j2⇒a2	91%
6	g2,j2⇒i1	91%
7	a2,b1⇒k2	83%
8	a2,f2⇒k2	77%

将获得的关联规则与具体的电信企业客户流失数据对应，达到根据客户流失数据给电信企业提供辅助决策支持的目的，预防客户的流失。具体对应关联规则说明如下（只是关联规则中的一部分）：

规则 1：c1⇒a2。置信度为 100%。意思理解为如果客户所使用的服务有附属客户的时候则客户不是老年人。

规则 2：d2⇒a2。置信度为 100%。意思理解为如果客户使用公司服务的月数为高时候，则客户不是老年人。

规则 3：f2⇒a2。置信度为 100%。意思理解为如果客户没有办理多条电话服务渠道，则客户不是老年人。

规则 4：k2⇒a2。置信度为 100%。意思理解为如果客户流失，则客户是老年人。

规则 5：h2,j2⇒a2。置信度为 91%。意思理解为如果客户没有使用网络且客户每个月的支出为高则客户不是老年人。

规则 6：g2,j2⇒i1。置信度为 91%.意思理解为如果客户的网络服务提供线路是光纤线路，且客户的每个月支出为高，则客户的合约方式为每月签约。

规则 7：a2,b1⇒k2。置信度为 83%。意思理解为如果客户不是老年人且客户有合伙人则客户流失。

规则 8：a2,f2⇒k2。置信度为 77%。意思理解为如果客户不是老年人且客户没有办理多条服务渠道，则客户流失。

通过上述的规则，我们可以从中获得一些客户数据之间的规则，这些规则能够给电信公司提供辅助的决策支持。比如说通过规则 1 我们可以知道老年人不需要附属客户的服务，这样在电信公司进行服务推广的时候，就可以尽量减少老年人的推广，提高推广效率。通过规则 7 我们可以得出不是老年人且有办理合伙人服务的客户容易流失，因此电信公司可以根据这个规则给予这些客户更多的服务支持，挽留可能流失的客户。

5 结束语

概念格是形式分析的核心数据结构。同时概念格也是数据挖掘的一个强有力工具，本文主要介绍了概念格在客户流失数据分析中的应用，利用概念格的特点，在概念格生成的关联规则进行数据挖掘。通过对客户流失数据的数据挖掘给电信企业提供辅助决策的支持，预防客户的流失。

参 考 文 献

[1]WILLE R.Restructuring lattice theory:an approach based on hierarchies of concepts[C].Ordered Sets,1982:445-470.

[2]覃丽珍,李金海,王扬扬.基于概念格的知识发现及其在高校就业数据分析中的应用[J].山东大学学报(理学版),2015,50(12):58-64.

[3]陈朝晖. 基于概念格的数据挖掘研究及应用[D].西安电子科技大学,2014.

[4]魏玲,曹丽,祁建军,张文修.形式概念分析中的概

念约简与概念特征[J/OL].中国科学:信息科学 :1-17[2020-11-22].http://kns.cnki.net/kcms/detail/11.5846.TP.20201118.1153.002.html.
[5]谢志鹏,刘宗田.概念格与关联规则发现[J].计算机研究与发展,2000(12):1415-1421.
[6]杨霁琳.一种基于概念格的规则提取方法及其应用[J].计算机科学,2012,39(S3):204-206.
[7]李冠宇.智能信息处理课件:形式概念分析_第 3 章概念格[R].大连海事大学,2020.