



基于预测度量值的 IPTV 用户行为规则预测算法

汪敏娟^{1,2}, 嵇正鹏³, 吕超^{1,2}

(1.江苏省公用信息有限公司,江苏 南京 210006;

2.中国电信股份有限公司智慧家庭运营中心,江苏 南京 210006;

3.中国电信股份有限公司江苏分公司,江苏 南京 210003)

摘要:提出了一种符合用户行为的,基于海量 IPTV 用户特征数据,对 IPTV 用户进行分群和规则提取的算法模型。首先提出了符合用户点播使用行为的 IPTV 用户分群的描述维度,即通过基础属性描述用户分群、通过点播行为描述用户分群变化趋势。然后提出了预测度量值的概念,对用户分群的稳定性进行描述,并提出了对稳定的用户分群提取点播行为概率的算法。最后通过大量的 IPTV 运营数据对算法模型进行了验证分析。

关键词: IPTV;点播行为;等价类划分;信息熵;预测度量值;规则提取

中图分类号: TP181

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016153

A rules extraction algorithm for IPTV customers forecasting based on the forecasting entropy measurement

WANG Minjuan^{1,2}, JI Zhengpeng³, LV Chao^{1,2}

1. Jiangsu Public Information Co., Ltd., Nanjing 210006, China

2. Smart Home Operation Center of China Telecom Co., Ltd., Nanjing 210006, China

3. Jiangsu Branch of China Telecom Co., Ltd., Nanjing 210003, China

Abstract: An algorithm model conformed to the user behavior, based on the massive IPTV user characteristic data which extract rules and classify IPTV users was proposed. First, IPTV user group description dimension in accordance with the user on demand was put forward. Namely, the user group could be described by basic property and trend of user behavior could be described by users' demand behavior. Then the concept of prediction measurement was put forward, the stability of user group was described, and an algorithm which extracted demand behavior probability on stable user group was proposed. At last, the algorithm model was verified and analyzed by massive IPTV operation data.

Key words: IPTV, demand behavior, equivalent class, information entropy, prediction measurement, rule extraction

1 引言

IPTV (internet protocol television, 网络协议电视) 是以网络协议为基础, 面向电视终端, 通过宽带网向用户提供

交互式电视及增值服务的业务^[1,2]。在业务的发展和运营过程中, IPTV 平台已积累了海量用户特征属性数据。这些数据是在用户办理业务和使用业务过程中积累的、对用户基本信息和使用信息的描述。用户特征属性数据蕴含了丰富

的用户行为信息。如何在这些海量的信息中提取出有意义的用户行为特征,用于有效地指导 IPTV 业务运营与服务,已成为当前研究的热点问题。用户行为特征的预测算法研究,无论对于运营商,还是对于用户,都具有很高的经济价值和实际意义。

国内外很多学者、运营机构对于 IPTV 用户行为特征预测进行了研究。目前的研究工作主要通过计算用户与内容的关联关系、学习用户兴趣标签、用户聚类角度展开。参考文献[3]基于用户聚类的关联推荐算法,利用皮尔曼相关系数和斯皮尔曼相关系数来确定 IPTV 用户群与内容之间的偏好关系;参考文献[4]提出从业务的用户使用场景出发,结合用户标签加强推荐的针对性;参考文献[5]将一种单蚁群聚类应用于 IPTV 用户群偏好分析中,解决了用户偏好发现问题;参考文献[6]提出了在没有先验知识的情况下,利用生态位理论计算 IPTV 服务之间的“竞争系数”,从而选取更为符合期望的服务的方法。上述工作都对利用 IPTV 用户特征属性信息进行用户分类、通过用户分类进行行为预测提供了有效解决方式。

在 IPTV 的实际运营过程中发现,用户的特征属性数据可分为基础属性和点播行为两类。基础属性是描述用户基本特征的数据(如所在地区、职业、产品套餐类型等);点播行为是在运营过程中产生的,随着用户使用业务而动态变化的数据(如平均消费值、最常收看的频道类型等)。

用户的点播行为是从无到有、不断变化的,而往往具有相同基础属性的用户在点播行为上表现出了一定的共性。另外,由于用户使用 IPTV 业务均经历了从陌生到熟悉的使用习惯培养过程,故处于不同状态阶段的用户特征属性数据对行为特征预测的贡献是不同的。可见应该从已趋于稳定状态的用户群体中获取行为特征,这样得到的规则较为贴近实际运营情况。而目前从上述角度进行的工作较少。

为此,从研究某些 IPTV 用户群是否适宜于进行用户行为预测入手,建立了一种全新的 IPTV 用户行为特征预测算法模型。首先利用 IPTV 用户的基础属性对用户进行基本分类,而后对每个基本分类中用户的点播行为是否已经趋于稳定进行观测。对于用户点播行为已趋于稳定的用户进行基本分类,计算出分类中点播行为出现的概率数据,利用此概率数据对 IPTV 用户的行为进行预测。通过对实际运营应用数据的分析与比较,提出的算法模型有

效地提升了 IPTV 用户行为预测的准确性和用户行为预测的计算效率。

2 IPTV 用户基本分类与预测度量值

2.1 IPTV 用户基本分类模型

研究的 IPTV 用户行为特征预测算法模型,是以用户特征属性信息为基础进行知识学习的计算模型。首先基于对用户在各基础属性上的取值,通过等价关系运算来对 IPTV 用户进行等价划分,每个等价划分对应一个用户基本分类。

设 IPTV 用户集合为 $U=\{u_1, u_2, \dots, u_n\}$ 。设集合 T 为用户的基础属性集, D 为用户点播行为集。 T, D 均为元素有限的已知集合,且 $T \cap D = \emptyset$ 。

设 $f_j: U \rightarrow V_j, (1 \leq j \leq m)$ 为用户属性取值判定函数,则 $f(u_i, t_j) = v_{ij}$ 表示用户 u_i 在属性 t_j 下的属性取值为 v_{ij} , $f(U, t_j) = V_j$ 表示用户集合 U 在属性 t_j 上的取值集合。

设关系 R_T 表示两个用户在基础属性集 T 上等价,即 $\exists u_i \in U, u_j \in U (1 \leq i \leq n, 1 \leq j \leq n, \text{且 } i \neq j)$, 使 $\forall t_k \in T, f(u_i, t_k) = f(u_j, t_k)$ 。

定义 1 集合 $X_i = \{u_j | (u_i, u_j) \in R_i\}$ 为 IPTV 用户群的一个基本分类。

即在 IPTV 用户集合 U 中,存在若干独立的用户,其在用户基础属性集 T 上等价。那么通过关系 R_T 可将 IPTV 用户分为若干基本分类: $\{X_1, X_2, \dots, X_m\}, (m \leq n)$ 。

基于上述定义对 IPTV 用户进行基础分类划分的算法模型如图 1 所示。通过对 IPTV 用户核心基础属性值与已知各 IPTV 用户基本分类在核心基础属性上的取值相比较,在取值函数上取值相同的,即可将用户划分到该基本分类中。不断地重复上述过程,即可不断地根据用户基础属性的取值来建立用户基本分类。

2.2 基本分类的预测度量值

通过把 IPTV 用户划分为若干个基本分类后,即可开始对各分类在运营过程中的变化情况进行观测。在参考文献[7,8]中提出,在 IPTV 运营过程中,绝大多数的用户会经历从对业务不熟悉到逐渐形成自己使用习惯的过程。即大多数用户使用了一定时期 IPTV 业务后,其对业务的使用习惯一般会停留在一个稳定的、有规律的状态附近。但如何来确定某一个用户基本分类已经趋于稳定,我们借鉴 Shannon 等人在参考文献[9-11]提出的经典信息熵概念:任何信息都存在冗余,冗余的大小与信息微观的不确定性有关。而信息

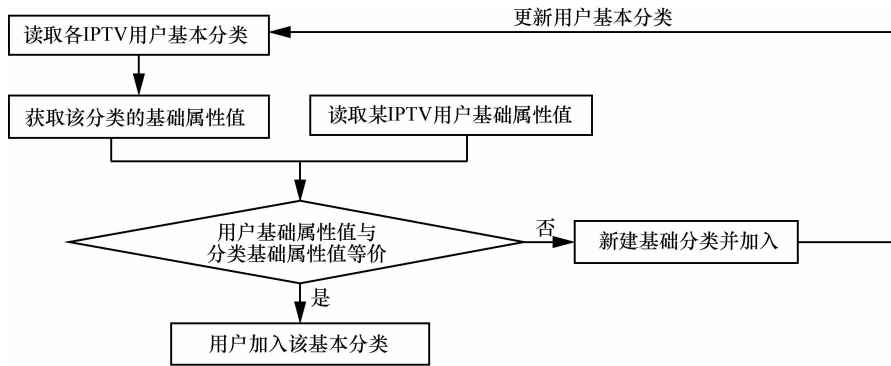


图1 IPTV用户基本分类划分算法模型

熵就是用来描述信息源不确定程度的概念,即熵值越大,其不确定性越大;熵值越小,其确定性越大。

在对IPTV用户的预测过程中,对于已经趋于稳定的用户基本分类,其用户体现同类型点播行为的可能性越大。本文提出的预测度量值,就是利用信息熵的概念,对每个基本分类中海量的用户点播行为进行概率计算,用于判定当前的用户基本分类是否已经趋于确定。

设通过第2.1节的理论对IPTV用户集 U 划分得到的用户基本分类为 $\{X_1, X_2, \dots, X_n\}$ 。在某一个基本分类 X_i 中,设其点播行为集 $D=\{d_1, d_2, \dots, d_l\}$,由于点播行为是变化的,故在基本分类中,每个用户在相同的点播行为上取值可能不同。设基本分类 X_i 中所有用户在属性 d_j 上的取值组成取值集合 $V_j=\{v_{j1}, v_{j2}, \dots, v_{js}\}$ 。设 $P(v_{jk})$ 表示在基本分类 X_i 中,点播行为 d_j 的取值为 v_{jk} 的概率。

依据信息熵的计算法则,定义IPTV用户基本分类点播行为的信息熵如下。

定义2 某IPTV用户基本中分类点播行为 d_j 的信息熵为:

$$H(d_j) = - \sum_{k=1}^s P(v_{jk}) \times \lg P(v_{jk}), v_{jk} \in V_j, 1 \leq k \leq s \quad (1)$$

$H(d_j)$ 描述了某一个点播行为的不确定性。它通过其所有可能取值的变化情况来描述该属性当前的状态。对于一个IPTV用户基本分类来说,其预测度量值为其包含的所有点播行为信息熵之和。

IPTV用户基本分类的预测度量值为:

$$H(X_i) = \sum_{j=1}^l H(d_j) \quad (2)$$

$H(X_i)$ 在传统的对信息变化描述的基础上,通过对IPTV用户基本分类中点播行为变化情况的汇总,来表示当前分类整体点播行为的不确定性。

使用预测度量值表示在一个IPTV用户的基本分类中,其所有用户的点播行为是否已趋于稳定状态。预测度量值越大,表示该分类中的用户行为仍处于较为混乱的状态,意味着该分类暂时无法用于用户预测;预测度量值越小,表示该分类中的用户行为已趋于稳定,意味着该分类已形成了一类用户行为的描述。即随着用户逐步形成自己的使用习惯后,其点播行为虽仍在不断变化,但已趋于某一种规律。

IPTV用户基本分类预测度量值算法具体如下。

Function Get Entropy_Measurement

for each X_i in $\{X_1, X_2, \dots, X_n\}$

var H_i ;

for each d_j in $\{d_1, d_2, \dots, d_l\}$

var temp_H;

for each v_{jk} in $\{v_{j1}, v_{j2}, \dots, v_{js}\}$

$$P(v_{jk}) = \frac{\text{count}(v_{jk})}{|X_i|};$$

temp_H += $P(v_{jk}) * \lg P(v_{jk})$;

end for v_{jk}

$H_i += 0 - \text{temp_H}$;

end for

let $H_i = H_i$;

put H_i into $\{H_1, H_2, \dots, H_n\}$;

end for

在算法伪代码中, $\text{count}(v_{jk})$ 表示在点播行为 d_j 的所有取值中, v_{jk} 出现的次数。 $|X_i|$ 表示集合 X_i 的势。

3 IPTV用户行为规则分析

得到每个IPTV基本分类对应的预测度量值后,对低于预设阈值的基本分类,认为其已基本趋于稳定。在IPTV

业务运营过程中,可根据当前各基本分类预测度量值实际取值和运营经验来预设阈值。一般来说,在一定时间周期内,某个用户基本分类的预测度量值停留在一个较低的取值区间且没有发生较大范围的波动时,可认为其已趋于稳定。

在趋于稳定的用户基本分类 X_i 中,用户在基础属性上等价,在点播行为上取值不同,但保持了一定的变化规律。因此可通过对基本分类 X_i 中 IPTV 用户基础属性取值与点播行为取值的统计关系来得到某一类 IPTV 用户的行为特征。

设点播行为集 $D=\{d_1,d_2,\cdots,d_l\},\forall d_j\in D$ 的取值集合 $f(X_i,d_j)=\{v_{j1},v_{j2},\cdots,v_{js}\}$ 。

则:

$$P(d_j=v_{jk})=\frac{count(v_{jk})}{|X_i|} \tag{3}$$

其中, $count(v_{jk})$ 表示在用户基本分类 X_i 中用户点播行为 d_j 取值为 v_{jk} 的数量; $|X_i|$ 表示用户基本分类 X_i 的势,即基本分类 X_i 中包含的用户数量。

上述规则可按如下方式进行描述:在基本分类 X_i 中,有 P 的概率会发生 $d_j=v_{jk}$ 的事件。

IPTV 用户基本分类的点播行为预测算法如下:

```
Function Forecasting
  for each  $X_i$  in  $\{X_1,X_2,\cdots,X_m\}$ 
    for each  $d_j$  in  $\{d_1,d_2,\cdots,d_l\}$ 
      for each  $v_{jk}$  in  $f(X_i,d_j)$ 
         $P=\frac{count(v_{jk})}{|X_i|}$ 

        There would be  $(P\times 100)\%$  possibility that
         $X_i$  would do  $d_j$ 
      end for
    end for
  End Function
```

4 实验分析

对 N 市 12 个月内 IPTV 用户的行为进行了预测实验。数据包括了 N 市 IPTV 用户的基础属性和第 1-12 个月的点播行为数据。具体数据见表 1。

首先根据用户的基础属性对用户进行分类,计算其预测度量值,并通过对已趋于稳定的用户基本分类进行用户行为特征预测,并将预测结果与 12 月的实际点播行为情

表 1 IPTV 用户特征属性集

属性分类	属性名称
基础属性	年龄段
	家庭类型
	职业
	学历
	用户类型
	地址(区)
点播行为	所属业务分组
	月平均消费值
	最常观看直播频道
	最常观看点播标签

况进行比对,以验证预测规则的准确性。

按照提出的用户基本分类算法,将 N 市分成了 30 余个用户基本分类,针对各分类计算其 1-11 月的预测度量值。表 2 是 N 市用户基本分类取值范围的情况。从表 2 中可以看到,用户基本分类的预测度量值处于一个“两头小、中间大”的状态,即部分基本分类已趋于稳定、少量分类处于大量变化阶段、多数用户分类处于由不稳定到稳定的变化过程中。

表 2 N 市用户基本分类 11 月预测度量值情况

预测度量值取值范围	用户基本分类个数
[0,5)	3
[5,10)	5
[10,15)	8
[15,20)	9
[25,30)	12
[35,40)	1

选取了预测度量值最小的 3 个基本分类、预测度量值最大的 1 个基本分类。其预测度量值 1-11 月份的变化情况如图 2 所示。从图 2 中可以看到,通过 11 个月的运营,用户基本分类 A、B、C 的预测度量值已逐步由较高的取值降低到一个平缓的取值范围,但也存在用户基本分类 D,其预测度量值始终处于不断变化的状态。

通过对上述用户基本分类的走访,发现用户基本分类 D 中的绝大多数用户属于集体宿舍、宾馆等形态的用户。用户流动性大、使用人群不固定,故其预测度量值始终处于一个变化的范围。而对于基本分类 A、B、C 来说,其绝大多数 IPTV 业务用户为普通家庭用户,通过一段时间的业务使用,其使用习惯已逐渐趋于稳定,月均消费额、最常观看的直播频道、点播标签等运营特征数据均已处于一个稳

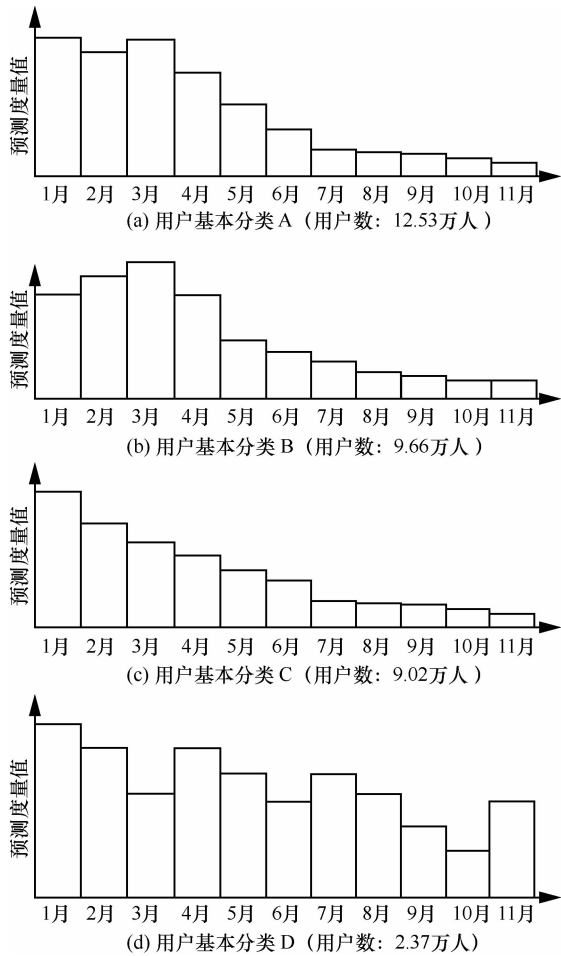


图2 N市部分IPTV用户基本分类预测度量值变化情况

定的范围内。

对于已趋于稳定的用户基本分类A、B、C,可通过其特征属性进行用户行为特征的预测。通过IPTV用户的预测算法,提取用户基本分类A、B、C中的行为规则,并将预测的规则与这3个基本分类在12个月的实际点播行为进行对比,用于验证预测规则的准确性。对比情况

见表3。

从表3可以看到,3个用户基本分类点播行为的预测值和实际值的误差率都在8%以内,能够有效地为运营人员针对不同的用户基本分类进行有针对性的运营提供数据基础。

5 结束语

提出了一种符合IPTV业务特征的分类及行为分析算法,结合实际运营经验将IPTV用户特征属性划分为基础属性和点播行为,利用基础属性对用户进行分类,利用点播行为描述各用户分类的变化情况,当用户分类变化趋于稳定时,计算该基本分类中用户的基础属性与点播行为之间的概率关系,并将该概率关系输出为行为规则,用于指导IPTV业务运营。通过对实际运营数据的分析,本文提出的预测算法明显地提升了IPTV用户预测的准确性,也降低了IPTV用户分类、行为规则提取的计算开销。

在后续的工作中,需对IPTV用户的特征属性分类提出更为精确的划分标准,并优化对IPTV用户基础属性分类算法和计算规则的约定,进而通过大规模数据统计规律来优化对用户点播行为信息熵、预测度量值的阈值估算方法。

参考文献:

- [1] 韦乐平. 三网融合与IPTV的发展和挑战[J]. 电信科学, 2006, 22(7): 1-5.
WEI L P. Triple-play and the development and challenges of IPTV[J]. Telecommunications Science, 2006, 22(7): 1-5.
- [2] XIAO Y, DU X, ZHANG J, HU F, et al. Internet protocol television (IPTV): the killer application for the next-generation internet [J]. IEEE Communications Magazine, 2007, 45(11): 126-134.

表3 预测规则与实际规则对比分析

用户基本分类	点播行为	预测值	实际值	误差率
A	月平均消费额/元	26.4	25.73	2.6%
	最多观看直播频道	体育频道,53.3%	体育频道,56.07%	-4.94%
	最多观看点播标签	纪实,42.26%	纪实,39.9%	5.91%
B	月平均消费额/元	35.52	37.08	-4.21%
	最常观看直播频道	娱乐频道,61.8%	娱乐频道,59.41%	3.87%
	最常观看点播标签	偶像剧,77%	偶像剧,82.9%	-7.12%
C	月平均消费额/元	7.03	7.49	-6.14%
	最常观看直播频道	城市频道,35.72%	城市频道,38.6%	-7.46%
	最常观看点播标签	养生,79.65%	养生,82.8%	-3.95%

- [3] KIM M W, SONG W M, SONG S Y, et al. Convergence and hybrid information technology [M]. Berlin: Springer, 2012: 409-416.
- [4] 朱映波, 刁建伟, 康波. 基于标签对家庭 IPTV 业务个性化推送机制的用户体验优化研究 [J]. 电信科学, 2014, 30(7): 113-120.
- ZHU Y B, DIAO J W, KANG B. User experience optimization researchbased on label of IPTV households personalized pushmechanism [J]. Telecommunications Science, 2014, 30(7): 113-120.
- [5] 曹三省, 孟静, 杜怀昌, 等. 蚁群聚类算法在 IPTV 用户群偏好分析中的应用[J]. 中国传媒大学学报(自然科学版), 2009, 16(1): 33-37.
- CAO S X, MENG J, DU H C, et al. Ant clustering algorithm in the application of IPTV user group preference analysis [J]. Journal of Communication University of China (Natural Science Edition), 2009, 16(1): 33-37.
- [6] SEOL H, PARK G, LEE H. Demand forecasting for new media services with consideration of competitiverelationships using the competitive bass model and the theory of the niche [J]. Technological Forecasting & Social Change, 2012, 79 (7): 1217-1228.
- [7] HELD G. Understanding IPTV[M]. NewYork: ACM Press, 2006.
- [8] JIM J. Analysis of the IPTV increment service development strategy[J]. Science & Technology Innovation Herald, 2014.
- [9] SHANNON C E. A mathematical theory of communication: mathematical theory of computation[M]. New York: McGraw-Hill, 1974: 3-55.
- [10] LIANG J, SHI Z, LI D, et al. Information entropy, rough entropy and knowledge granulation in incomplete information systems[J].

International Journal of General Systems, 2006, 35(6): 641-654.

- [11] VALLÉ R. Information entropy and state observation of a dynamical system [C]// International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, June 20-24, 1987, Eindhoven, Netherlands. Berlin: Springer-Verlag, 1987: 403-405.

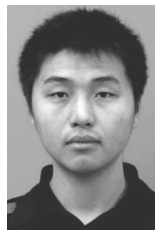
[作者简介]



汪敏娟(1975-),女,江苏省公用信息有限公司副总经理,中国电信股份有限公司智慧家庭运营中心副主任、高级工程师,主要研究方向为 IPTV 运营规律、视频承载网络规划、知识学习。



嵇正鹏(1964-),男,中国电信股份有限公司江苏分公司企业发展部主任、教授级高级工程师,享受国务院特殊津贴,主要研究方向为多媒体网通信技术、视频运营、数据挖掘。



吕超(1985-),男,江苏省公用信息有限公司、中国电信股份有限公司智慧家庭运营中心工程师,主要从事大数据平台开发、数据运营与维护工作。