文章编号：1003-0077(2016)04-0213-10

在线游戏用户的流失预测：基于不平衡数据的采样方法比较和分析

吴悦昕，赵 鑫，过岩巍，闫宏飞

（北京大学 计算机科学技术系，北京 100871）

摘 要：流失用户预测问题在很多领域都是研究重点。目前主流的流失用户预测方法是使用分类法，即把用户是否会流失看作一个二分类问题来处理。该文提出了一个基于二分类问题解决的在线游戏流失用户预测方法。此方法除了总结了一些对在线游戏而言比较重要的可以用于流失预测的特征之外，也考虑到流失用户相对稀少的问题，在流失用户预测问题中引入了不平衡数据分类的思想。该文主要在流失预测中结合使用了基于采样法的不平衡数据处理策略，并对现有主要的几种采样算法进行了对比实验和分析。

关键词：在线游戏；流失预测；不平衡数据；采样法

中图分类号：TP391 **文献标识码**：A

User Churn Prediction for Online Game: Comparison and Analysis of Approaches Based on Sampling for Imbalanced Data

WU Yuexin, ZHAO Xin, GUO Yanwei, YAN Hongfei

(Department of Computer Science and Technology, Peking University, Beijing 100871, China)

Abstract: The problem of user churn prediction is a research focus in many fields. Currently the main approach of the problem is based on classification, which predicts whether users will churn by a 2-class classification process. This paper addresses an approach for online game user churn prediction based on 2-class classification. We summarize some important features for the problem of online game user churn prediction. Furthermore, we noticed that churned users is relatively rare, and introduce the imbalanced learning methods into our work with a focus on the sampling methods. We conducted experiments on major sampling methods and analyzed the results.

Key words: online game; user churn prediction; imbalanced data; sampling

1 引言

流失用户预测问题是一个被广泛关注的重要而困难的问题。在电信[1]、银行[2]、电子商务[3]等领域，流失用户预测都是一个重要的研究方向。文献[1]表明，对于电信业来说，赢得一个新客户所花费的成本约为$300～600，这大约是保留一个老客户所花费成本的5～6倍。这对于在线游戏领域来说也是相似的。特别是目前主流的依靠对附加内容收费的在线游戏，尤其依赖于频繁、大量向游戏付费的高付费玩家。高付费玩家数占总玩家数的比例较小（在数万的总游戏人数中只有4 000个左右），因此吸引一个高付费玩家进入游戏的成本与保留一个老的高付费玩家的成本之比相对于每个用户都需要付费的电信业来说会更加高昂。这里，本文主要研究在线游戏内高付费用户的流失预测问题。相对于以前的研究，本文主要把重点放在了对数据的预处理上。因为流失预测问题的特点，流失的用户往往是十分少量的，而正常的活跃用户数量相对来说则过于庞大。这种数据的不平衡性大大影响了预测效果。本文通过预先对数据平衡化，使得预测结果的F值得到了15%以上的提升，效果十分明显。这说明数据平衡化是提升流失预测结果的一个简单、有效的手段。

收稿日期：2014-09-10 **定稿日期**：2015-03-15

基金项目：973项目(2014CB340400)；国家自然科学基金(61272340)；江苏未来网络创新研究院项目(BY2013095-4-02)

根据其他行业内的相关研究[1-2]，我们发现目前对此类问题主流的处理思路是将其看作二分类的问题，使用有监督的机器学习的方法来解决。根据这个思路，我们首先需要在游戏原始记录中总结出与用户流失相关的一些特征，然后利用已知流失与否的用户记录来训练并测试分类器，最后测试效果较好的分类器即可用于用户流失预警的任务。

实际应用过程中，我们发现流失的用户数量远远小于未流失的活跃用户数量的。我们手头的数据当中流失用户与没有流失的活跃用户数量之比约为1∶7，属于不平衡数据。传统的有监督的分类模型和算法都必须在相对平衡的数据上才能有比较好的效果，而在不平衡度较高的时候，则会对多数类别产生严重的偏向，有时候甚至会出现学习到的分类器会把所有输入的未标记数据都标记为多数类别的情况。不平衡数据问题在各个领域的流失用户预测问题中基本上都是普遍存在的，但目前对流失用户预测问题的研究方向主要是深入挖掘原始记录，对各个记录与用户的流失倾向性的关联进行分析（如文献[4]通过数据挖掘发现呼叫模式变化可以有效预测电信用户流失），以及使用复杂的分类模型，使之更适应于流失预测的任务（如文献[5]将混合过程神经网络方法应用到了流失用户预测任务中），并没有对不平衡数据进行针对性的处理。文献[6]考虑到了不平衡数据对预测的影响，但只采用了基于代价敏感学习的思路，通过改进的支持向量机来建立模型，方法比较单一，缺乏通用性。

而我们则尝试转换思路，使用了采样法对训练数据集进行调整，实现不平衡数据的平衡化处理。这样的处理方法通用性强，不需要过于深入挖掘特征和研究复杂模型，可以很容易地应用到不同领域的流失用户预测当中。本文研究了目前主流的基于采样的不平衡数据处理方法，将其结合到我们的流失用户预测问题上进行了实验，并对这些方法进行了测试、分析和比较。在不进行不平衡数据处理时，应用支持向量机进行分类实验只能达到32.8%的正类F值和0.600的正类ROC-AUC。在使用采样法进行处理之后，这两者最高分别被提升到48.7%和0.737，提升十分显著。

2 在线游戏流失用户预测问题和方法简介

2.1 问题和处理框架

我们的在线游戏流失用户预测任务的问题在于根据所有高付费用户最近一段时间的原始游戏记录来预测哪些用户会在一段较短时间内有较大可能性从游戏中流失。

因为我们已经拥有了所有高付费用户的完整游戏记录以及他们的流失情况，因此我们可以使用有监督的机器学习方法来寻找这些游戏记录与用户的流失倾向之间的内在关系。我们使用二分类问题的框架来处理流失预测问题：每个用户有一个表示其是否流失的流失标签以及一系列状态特征，对于有确定标签的用户，我们使用其状态特征来训练一个两输出的分类器用于预测无标签用户的流失标签。由于我们已经有了用户的流失标签，因此我们的任务在于以下两方面：从庞杂的游戏记录中总结出与用户流失倾向相关的状态特征，以及找到一个能够尽可能提升预测结果的分类器训练方法。

2.2 特征提取

我们找到的与在线游戏流失用户预测任务相关的论文只有文献[7]，而此文献使用基于社会影响的方法进行分析，这与我们通过游戏行为分析的任务不符。由于没有相关的工作可以参考，因此我们自行对游戏记录进行了一定的分析，提取了一些特征。我们希望提取的特征可以在计算上比较简单，与我们手头的游戏记录能够比较契合，并且能够大致对用户的活跃程度进行描述。

我们最后使用的特征有四大类，共17小类，具体的每类特征选用见表1。对每个小类的特征来说，具体的特征以天为单位计算。以登录时间为例，我们计算用户每天的登录时间，并将其作为一个特征，故每小类特征中的特征数等于我们考虑的游戏天数。对每个用户来说，我们选取其最后一次登录之前若干天的游戏情况作为其特征。例如，使用十天的游戏情况，则按之前的描述，就会产生170个特征。这些特征都会进行归一化处理。

表1 特征列表

特征大类	特征名称	特征说明
在线情况	登录次数	用户每天登录游戏的次数
	在线时间	用户每天进行游戏的时间

续表

特征大类	特征名称	特征说明
货币花费	货币花费总量	用户每天花费的真实货币总量
	真实货币换取游戏币花费	用户每天花费真实货币换取的游戏货币占所有获得的游戏货币之比
互动情况	军团操作数	用户每天执行的军团操作数量
	组队操作数	用户每天执行的组队操作数量
	国战参与数	用户每天参加国战次数
	国战参与所获	用户每天参与国战获得的威望值
	攻击其他玩家次数	用户每天攻击其他玩家次数
	攻击其他玩家奖励	用户每天攻击其他玩家获得的威望值
其他行为	武将操作数	用户每天进行的武将操作次数
	威望总量	用户每天的累计威望总量
	威望获得总量	用户每天获得的威望总量
	筑城参与数	用户每天参与筑城的次数
	筑城参与所获	用户每天参与筑城获得的提升
	征收事件数	用户每天遇到的征收事件次数
	征收事件所获	用户每天从征收世界获得的奖励数

最后，为了实现预测，每个用户最后若干天的游戏情况将不参与到特征计算中。例如，我们不考虑每个用户最后四天的游戏情况，意味着我们意图实现一个能够至少提前四天预测用户是否流失的分类器。

2.3　分类器训练与不平衡数据

确定需要使用的状态特征之后，我们可以把每个用户表示为一个二元组(x,y)，其中x为我们选用的状态特征组成的特征向量，y为类别标签（流失或活跃）。给定一组用户数据集合$\{(x,y)\}$，我们可以利用其训练一个分类器。训练得到的分类器可以用于预测无标签数据的类别，实现流失预测。这里我们定义流失用户为正类，活跃用户为负类。

通常，流失用户的数量大大低于活跃用户的数量。对于传统分类器来说，不平衡数据会对其性能产生显著的影响[8]。传统分类器在训练阶段并不考虑数据中可能的不平衡性，在构造一个对于训练数据集错误率最小的模型的时候，就会产生对于多数类别的严重倾向。这是由于少数类的实例过于稀疏，使得分类器无法正确学习到其中的各个子概念[9-11]。对于多数类来说，由于拥有庞大的数据，这种没能被规则充分描述的子概念很少出现；而对于少数类别来说，这种情况就比较严重，分类器很难判断对于一些少数类实例，是应该视其表达了一个子概念，还是将其视为噪音。因此，这样学习到的模型无法对少数类有较好的分类效果。

鉴于传统分类器在大部分问题上的有效性，我们还是在应用传统分类器的基础上进行不平衡数据的处理，目前的研究也主要基于这个方向。一些方法只对某种特定的分类器有用，如决策树[12]和神经网络[13]，因此在应用上有不少局限。本文主要着眼于能与大部分分类器配合的具有一般性的方法。处理不平衡数据的主要思路是使数据平衡化，而数据平衡化可以在训练前或训练时完成。采样法[14-15]通过在训练前对数据平衡化来解决不平衡数据问题，而代价敏感学习[16]则采用的是在训练时对少数类进行补偿的方法。研究表明，代价敏感学习与以采样法有很强的相关性[17-19]，因此本文主要基于采样法来对用户流失预测问题进行处理。

3　使用不平衡数据进行用户流失预测

3.1　采样法概述

所谓采样法（Sampling），是一种处理数据的技术。其主要思路是对不平衡的训练集数据进行修改，构造出一个不平衡度减小的相对平衡的数据集。采样法主要分为两种，Under Sampling 与 Over Sampling。本文定义所有用于训练的已经有标签的用户特征数据构成集合S，S_{maj}为S中所有活跃用户的集合，S_{min}为S中所有流失用户的集合。顾名思义，Under Sampling 方法减少S_{maj}中的用户数，得到其的一个子集E_{maj}，并让其与S_{min}一同训练分类器。

Over Sampling 方法则相反，通过增加 S_{min} 的用户数，得到新集合 E_{min}，然后让其与 S_{maj} 一同训练分类器。

假定我们手头的数据中有 500 个活跃用户，50 个流失用户。直接使用这些数据训练分类器得不到很好的效果，于是我们事先对数据进行采样处理。如果我们选择使用某种方法将活跃用户数量减少，假设减少到 100 个，这就属于 Under Sampling 方法；如果我们选择某种手段将流失用户数量增加，假定增加到 300，这就属于 Over Sampling 方法。

下面介绍几种常用的采样算法。

3.2 随机采样

随机采样分为随机 Under Sampling 与随机 Over Sampling。随机 Under Sampling 就是说从 S_{maj} 中随机选出一个事先给定了大小的子集构成集和 E_{maj} 来代替 S_{maj}。而随机 Over Sampling 则不断随机从 S_{min} 中选取用户，然后将其副本放入 S_{min}，直到其成为一个事先给定了大小的集合 E_{min}，并用其替代原来的 S_{min}。这两种算法的优点是简单，容易理解和实现。

如果参照我们上面的例子，随机 Under Sampling 算法会随机从 500 个活跃用户中选择 100 个用于最终训练，而随机 Over Sampling 算法会随机创建流失用户数据的副本直到数量达到 300，然后进行训练。

两种随机采样方法看起来是等价的，因为他们可以把原数据集调整到一个相同的不平衡度。但实际上，两者都有各自的问题，使得分类器学习到的模型产生偏误[10,20-21]。随机 Under Sampling 的问题比较明显，就是可能会把 S_{maj} 中体现活跃用户概念的较重要、信息量大的用户移除，降低分类器的学习效果[22]。随机 Over Sampling 的问题则比较隐蔽。其问题在于，随机 Over Sampling 的过程相当于产生 S_{min} 中用户的简单拷贝，因此在特征空间中某些点会堆积过多的用户实例，使得分类器的训练产生过拟合的现象，即训练得到的模型过于复杂使得能够比较精确地拟合训练集中的用户，但对新用户的分类效果却产生了下降[20]。

3.3 有导向的 Under Sampling

随机 Under Sampling 的问题是可能会移除比较重要的用户，因此改进的方法就是分析 S_{maj} 中的用户特征，并移除其中相对不重要的那些用户，达到 Under Sampling 的效果。这就形成了有导向的 Under Sampling 方法。

一种检测用户信息的方法是使用用户特征的 K 近邻信息(KNN Under Sampling)[23]。此方法认为离 S_{min} 中用户距离较远(即与流失用户较不相似)的用户所含信息较少，并选取那些离 S_{min} 中用户距离较近的 S_{maj} 中用户来构成集合 E_{maj}。一个效果相对较好的距离计算方法是计算 S_{maj} 中每个用户与所有 S_{min} 中用户距离值当中 K 个最大值的平均值来作为其与 S_{min} 的距离。然后根据事先给定的数量选取距离较小的一部分用户组成 E_{maj}。以之前的例子来说，我们需要计算所有 500 个活跃用户和与之距离最远的 K 个流失用户的平均距离，然后选出此距离值最小的 100 个活跃用户用于最终训练。

另一种移除信息量小的用户的方法是利用所谓的浓缩近邻法(Condensed Nearest Neighbor Rule，简称 CNN)[23]。这个方法选取 S 的一个一致子集合 E 来代替 S。所谓 E 是 S 的一致子集合指 E 是 S 的子集且利用 E 训练的 1-近邻分类器可以对 S 进行完全正确的分类，即对 S 中每个用户找到其在 E 中距离最近的用户，两者所属类别相同。S 的一致子集合 E 的构造方法为，先取 E 等于 S_{min}，然后在 E 中加入任取的一个 S_{maj} 中用户。之后利用 E 对 S_{maj} 中每个用户进行 1-近邻分类，如果分类错误就把该用户加入 E。这样构造的一致子集合并不一定是最小的，但实践表明通过这个方法可以充分缩小原始数据集。CNN 方法通常会和之后提到的数据清理算法结合使用。

3.4 人工数据构造法

人工数据构造法是一种 Over Sampling 方法。由于随机 Over Sampling 方法容易产生过拟合的现象，为了减小过拟合，Over Sampling 方法加入的数据最好不是已有数据的简单拷贝。于是产生了人工数据构造法，将基于原数据集中用户构造的人工数据加入以实现 Over Sampling。

一个广泛使用的人工数据构造法是 SMOTE (the synthetic minority oversampling technique)[25]，是一个基于 K 近邻用户来构造人工数据的方法。SMOTE 方法为 S_{min} 中每个用户构造若干新用户。为 S_{min} 中用户 x_i 构造新用户时，先找到其在 S_{min} 中的 K 个最邻近用户，并在其中随机选取一个用户 x_j，则构造的新用户为 $x_{new}=x_i+(x_j-x_i)*\delta$，其中 δ 是 0 到 1 之间的一个随机数。实际上，构造的新

用户就是 x_i 与 x_j 在特征空间中连线上的一点。以前文的例子来说，我们需要构建 250 个人工流失用户。构造每个人工流失用户时，我们首先随机选取一个流失用户作为样本，然后再随机从它的 K 近邻中选取一个流失用户作为参考，新生成的流失用户是这两个流失用户连线上随机选取的一点。

3.5　有导向的人工数据构造法

SMOTE 方法构造人工数据时，S_{min} 中的每个用户的地位是相同的，根据每个用户构造的新用户数量是相同的。但实际上，每个用户的信息量不同，因此需要构造的人工用户的数量也往往不同。因此产生了根据用户的 K 邻近信息来计算需要生成的新用户数量的方法。

BorderLine 方法[26]只为 S_{min} 中“危险”的用户构造人工用户。所谓“危险”的用户指这样的用户，其在所有用户集 S 中的 K 近邻中，属于 S_{maj} 的用户数量大于等于 K/2 而小于 K。这里 K 近邻用户都属于 S_{maj} 时则被考虑为噪音而不为其构造人工数据。BorderLine 方法通过增加两类边界处的流失用户数量来丰富流失用户的边界，使分类器偏向流失用户。

ADASYN 方法[27]则比较直观。此方法计算 S_{min} 中所有用户的 K 近邻中属于 S_{maj} 的用户所占的比例，然后以此比例值为权值来分配每个用户需要构造的新用户的数量。这样，越“危险”的用户会被构造越多的新用户，分类器就会给予其更多的偏向。

3.6　数据清理方法

数据清理方法是一种清除类间重叠的采样方法。常用的数据清理方法是基于 Tomek Link 的数据清理方法[28]。Tomek Link 指一个用户对$<x_i, x_j>$，其中 $x_i \in S_{maj}$，$x_j \in S_{min}$，并且不存在 $x_k \in S$，使得 $d(x_i, x_j) > d(x_i, x_k)$ 或 $d(x_i, x_j) > d(x_k, x_j)$，其中 $d(x, y)$ 指用户 x 与 y 特征向量之间的欧氏距离。容易知道，一个 Tomek Link 中的两个用户或是位于类边界的两侧，或是至少有一个是噪音。使用 Tomek Link 进行数据清理时，可以将其作为一种 Under Sampling 方法，去除每个 Tomek Link 中属于 S_{maj} 的用户。此时，通常将其和之前提到的 CNN 方法结合使用。One Side Selection 方法[29]就是这样的 Under Sampling 方法，此方法先通过 Tomek Link 对原数据集进行一次 Under Sampling，然后再使用 CNN 方法进行一次 Under Sampling。由于 Tomek Link 的计算比较耗时，因此也有人先采用 CNN 方法，然后再使用 Tomek Link 进行 Under Sampling(CNN＋Tomek Link 方法)[6]。也可以将 Tomek Link 作为对其他采样算法进行数据清理的方法，此时可以通过清除每个 Tomek Link 中的所有用户来实现数据清理。基于 SMOTE 的方法常和此类数据清理方法结合使用[22]。

3.7　采样法总结

表 2 对本文之前介绍的各个采样方法进行了简单总结。

表 2　基于采样法的不平衡数据处理方法

采样方向	方法名称	方法简介
Under Sampling	Random Under	随机移除部分活跃用户
	KNN	根据 KNN 信息移除部分活跃用户
	OSS(One Side Selection)	先计算所有 Tomek Link 并移除其中活跃用户，然后应用 CNN 方法
	CNN＋Tomek Link	先应用 CNN 方法，然后计算所有 Tomek Link 并移除其中活跃用户
Over Sampling	Random Over	随机选取流失用户制造副本
	SMOTE	利用每个流失用户的邻近用户来创造新的人造用户数据
	BorderLine	只为近邻多为活跃用户的流失用户创造新的人造用户数据
	ADASYN	根据近邻中活跃用户数量的多寡决定每个流失用户需要创造的人造用户数量
	Random Over＋ Tomek Link	应用 Random Over 方法后计算所有 Tomek Link 并移除其中所有用户
	SMOTEr＋ Tomek Link	应用 SMOTE 方法后计算所有 Tomek Link 并移除其中所有用户
	BorderLiner＋ Tomek Link	应用 BorderLine 方法后计算所有 Tomek Link 并移除其中所有用户
	ADASYNr＋ Tomek Link	应用 ADASYN 方法后计算所有 Tomek Link 并移除其中所有用户

4 实验设置

4.1 数据准备

我们已经有了原始的游戏记录、用户列表、用户标签以及要抽取的特征列表。我们要做的是得到能够用于输入分类器的代表每个用户的特征和标签的组合。因为特征是以天为单位计算的，因此我们需要先扫描记录，把需要计算特征的用户的游戏记录按天分割开，然后为每个选定的用户逐天计算各个特征。每个用户的特征值需要进行归一化才能在分类中有较好效果。归一化过程先计算所有高付费用户每个特征每天的平均值，然后计算用户每个特征每天的值与对应的平均值之比，将其作为最后使用的特征值。最后根据需要使用的特征以及天数，构造可以用于分类训练和测试的特征文件。最后得到的数据集中一共有3 898个用户实例，其中496个属于正类(流失用户)，3 402个属于负类(活跃用户)。

4.2 结果评价

本文使用支持向量机(使用RBF核函数)作为基本分类器，并采用五折交叉验证的方式对结果进行评价。通常在不平衡数据中，人们重点关注正类的分类效果，因此对正类的分类结果单独计算得到的准确率、召回率、F值、ROC曲线等将更适合于评价对不平衡数据的分类效果。下面对本文使用的评价指标进行详细介绍。

对于二分类问题而言，每个用户的分类结果可能有四种情况，如图1的困惑矩阵所示。因此，对于正类来说，其准确率的定义为TP/(TP+FP)，即分类器报告的正类用户中真正正类用户所占的比率；召回率的定义为TP/(TP+FN)，即分类器正确报告的正类用户占所有正类用户的比率。正类的F值就是正类的准确率和召回率的调和平均数。负类的准确率、召回率、F值也可以按类似方式定义。

		实例类别 T	实例类别 N
预测结果	T	TP	FP
预测结果	N	FN	TN

图1 困惑矩阵

ROC曲线[19,22]是分类结果中TP率和FP率的曲线。TP率的定义为TP/(TP+FN)，等于正类召回率；FP率的定义为FP/(FP+TN)，等于1-负类召回率。使用ROC曲线来作为分类器效果的评价标准时，多采用ROC曲线下方面积(简写为ROC-AUC)来作为数值化的标准。图2中，孤线的结果优于直线，ROC-AUC也更大。

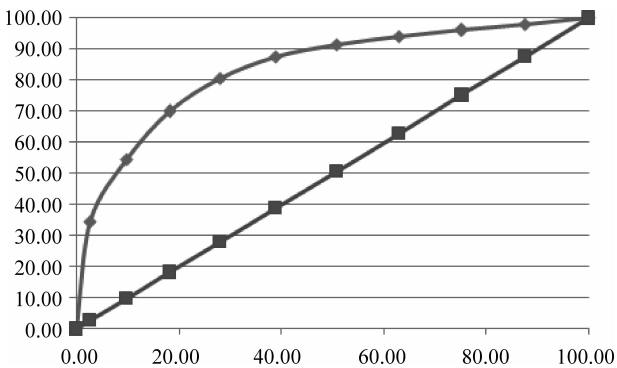

图2 ROC曲线示意

5 实验结果及分析

5.1 各种不平衡数据处理方法实验结果和分析对比

首先我们考察每种采样方法在设置不同的采样比率后可以达到的最好结果，评价标准分别为正类F值和ROC-AUC。此处我们设置使用所有特征，使用的特征天数为十天，提前天数为四天，使用五折交叉验证来检验分类结果。每个结果都是三次重复实验的平均值。

由图3可知，所有采样算法在最好情况下两个指标都大大优于不进行采样处理的情况。对于四种Under Sampling算法来说，两个指标在最好情况下都劣于所有的Over Sampling算法。Under Sampling算法中最好的是随机Under Sampling算法，说明其他方法在保留信息量大的负类用户方面效果都比较一般。Over Sampling算法相差都不大，其中最好的是ADASYN算法。另外在使用Tomek Link进行处理后，各个Over Sampling算法的效果都产生了一定的下降。这主要是因为本文使用支持向量机作为基本分类器，而支持向量机使用支持向量作为分类依据，因此对扩展类边界和移除噪音有帮助的数据清理算法对于支持向量机来说很难产生正向的改进。

然后我们来看采样比率变动对不同采样算法的影响。采样比率指应用采样法后被增加或减少的那

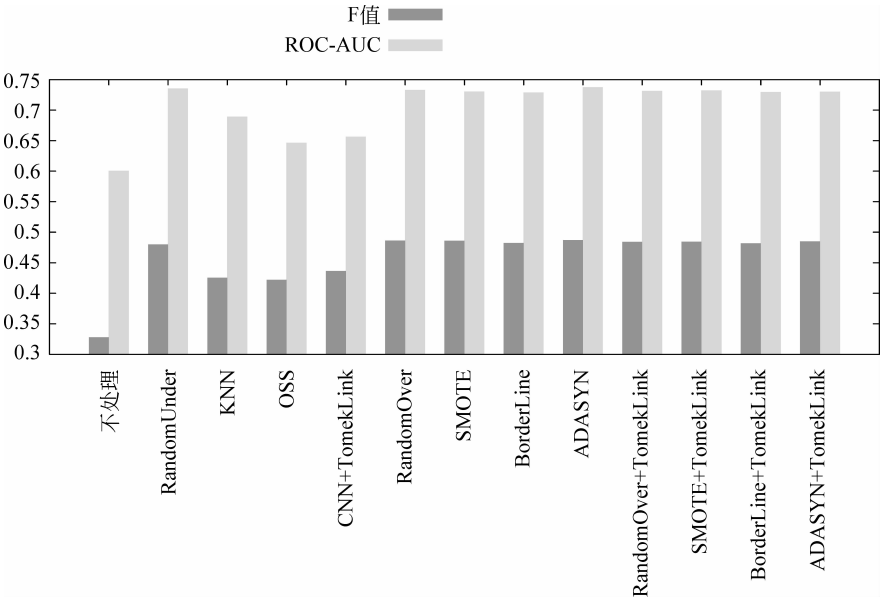

图 3　各采样法最佳效果对比

类的实例数与采样前之比。首先我们来看 Under Sampling 算法。图 4 左边是两种可改变采样比率的 Under Sampling 算法在不同比率下正类 F 值和 ROC-AUC 值的变化情况。可以看出几乎在所有比率下，两个指标都是随机 Under Sampling 算法较高。不过，对随机 Under Sampling 算法来说，两个指标有较明显的峰值，而 KNN 算法则相对平缓。然后考察采样比率变化时正类准确率与召回率的变化。图 4 右边表示，随着采样比率的升高，两者都出现正类准确率升高，而召回率下降的情况。这是因为采样比率提升的时候，数据集中属于负类的用户数量增加，此时分类器会缩小识别到的正类的概念空间，扩大负类的概念空间。在采样比率设置过低的时候，分类器学习到的正类的概念空间过大，因此会错误地把很多多数负类用户识别为正类，使得正类的准确率偏低而召回率较高。在采样比率提升时这种倾向就会逐渐降低，导致正类准确率升高，而召回率下降。另外，正类召回率基本上都是 KNN Under Sampling 较高。这似乎违反了直观，因为 KNN Under Sampling 优先保留与正类用户接近的负类用户，这样应该会减少识别到的正类概念空间，导致正类召回率降低。不过，事实上这些被优先保留的用户往往较多属于噪音而较少处于类边界上，而目前分类器对噪音都有一定的容忍度，因此识别到的正类概念空间在同等情况下会稍大。

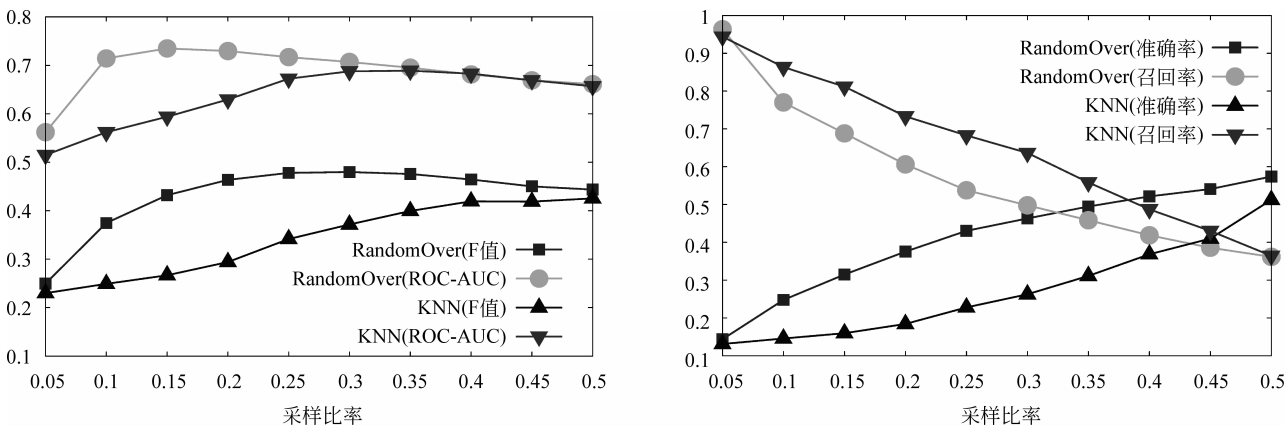

图 4　采样比率变化对 Under Sampling 算法的影响

对于 Over Sampling 算法来说，采样比率对准确率和召回率也有类似的影响。随着采样比率的升高，Over Sampling 算法出现正类准确率下降，而召回率升高的情况。这也是由于分类器学习到的正类的概念空间的改变所导致的，这里不再进行详细的分析。

5.2 改变使用的特征对结果的影响

这里我们考察使用不同大类的特征时对结果的影响。之前的结果都是使用了所有四大类特征得到的，这里尝试只使用其中的某个大类的特征来进行分类实验。结果如图5。由于实验结果当中F值与ROC-AUC的变化趋势基本相同，因此为了简明起见本文仅展示了ROC-AUC的结果进行对比。

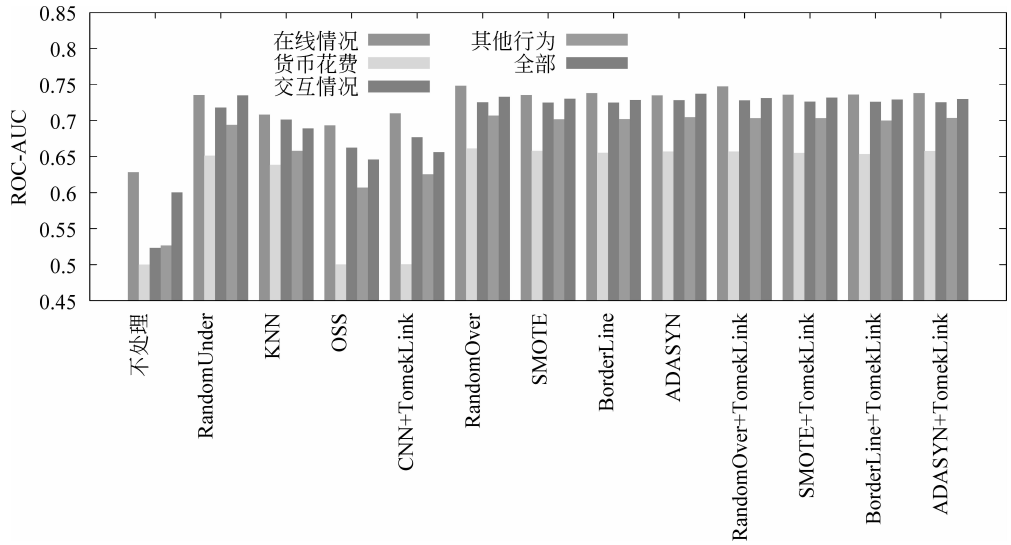

图5 单独使用不同类特征时对预测结果的影响

可以发现，单独使用在线情况进行分类实验的时候，在多数采样方法下结果是最好的，甚至优于使用所有大类特征时的效果。这说明在线情况在我们的数据当中是一个最有力的表现用户的活跃情况的特征。对于其他大类的特征来说，货币花费对于用户流失的预测效果是四大类中最差的。这点比较出乎我们的意料，因为我们预期对于高付费用户来说，真实货币的花费应是其活跃度的一个直接反映。实际分析之后发现，高付费用户在流失之前既可能如我们之前预测的那样减低货币花费，也可能反而增加花费。增加花费的一个可能是用户之前已经在游戏内充入一定量的真实货币，因此想在退出游戏之前将其消耗完；另一个可能是用户在离开游戏之前会有一定的赌博心态，从而会先执行一些充值抽奖类的操作，如果有好的获得则继续一段时间的游戏，否则彻底退出游戏。这样，货币花费对与用户活跃程度的预测能力就下降了。最后，我们发现在不进行采样处理时，单独使用除在线情况之外的某一大类特征时最后结果都较差，ROC-AUC与0.5十分接近。而进行采样算法后，效果大大提升，有些甚至已经比较接近使用所有大类特征时的效果。这表明在特征选取相对不完善时，不平衡数据会将这种不完善性放大，使得分类的结果急剧恶化。在使用采样法弱化不平衡数据问题之后，我们发现其实特征的不完善程度并没有太高，各个大类的特征都能够在一定程度上反映用户的活跃程度。也就是说，即使我们不能找到非常适合于流失预测的特征，在使用采样法之后我们也能取得相对可以接受的预测效果。

6 总结

在流失预测的任务当中，本文创新性地采用了基于采样法的不平衡数据处理方法，并将其应用在了一个新的领域——在线游戏领域中，取得了较好的效果。由于考虑了不平衡数据处理，因此即使在特征相对不完善时也能取得相对较好的预测效果。这样的结果为流失用户预测问题提供了一个新的思路，即在不过分深入地挖掘特征以及改进模型的情况下，通过对数据集的针对性处理来提升预测结果。未来我们可以继续尝试其他的不平衡数据处理法，以及将目前的方法应用到其他领域当中，通过继续研究来让我们的方法更加完善。另外，本文的方法主要还是一个离线算法，而实际的流失预测问题通常须要在一个在线的环境中实现动态地预测。为了将我们的方法应用到在线的环境中去，我们将来还需要考虑很多方面的问题，例如，对模型进行更新、重新训练的时机和如何加快训练、预测的速度等。这些也构成了未来流失预测问题研究方向的重要一环。

参考文献

[1] 夏国恩，金炜东. 基于支持向量机的客户流失预测模型 [J]. 系统工程理论与实践，2008，28(1)：71-77.

[2] 应维云，覃正，赵宇，等. SVM 方法及其在客户流失预测中的应用研究 [J]. 系统工程理论与实践，2007，27(7)：105-110.

[3] 朱帮助，张秋菊. 电子商务客户流失三阶段预测模型 [J]. 中国软科学，2010，(06)：186-192.

[4] Wei C P，Chiu I. Turning telecommunications call details to churn prediction：a data mining approach[J]. Expert systems with applications，2002，23(2)：103-112.

[5] Song Guojie，Yang Dongqing，Wu Ling，et al. A mixed process neural network and its application to churn prediction in mobile communications[C]//Proceedings of Sixth IEEE International Conference，2006：798-802.

[6] 钱苏丽，何建敏，王纯麟. 基于改进支持向量机的电信客户流失预测模型[J]. 管理科学，2007，20(1).

[7] Kawale J，Pal A，Srivastava J. Churn prediction in MMORPGs：A social influence based approach[C]//Proceedings of Computational Science and Engineering，2009. CSE′09. International Conference on. IEEE，2009，4：423-428.

[8] Chawla N V，Japkowicz N，Kotcz A. Editorial：special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations Newsletter，2004，6(1)：1-6.

[9] Weiss G M. Mining with rarity：a unifying framework [J]. Sigkdd Explorations，2004，6(1)：7-19.

[10] Holte R C，Acker L E，Porter B W. Concept learning and the problem of small disjuncts[C]//Proceedings of the Eleventh International Joint Conference on Artificial Intelligence. 1989，1.

[11] Quinlan J R. Induction of decision trees[J]. Machine learning，1986，1(1)：81-106.

[12] Maloof M A. Learning when data sets are imbalanced and when costs are unequal and unknown[C]//Proceedings of ICML-2003 workshop on learning from imbalanced data sets II. 2003.

[13] Hykin S. Neural networks：A comprehensive foundation[J]. Prentice Hall International，Inc，1999.

[14] Laurikkala J. Improving identification of difficult small classes by balancing class distribution[J]. Artificial Intelligence in Medicine，2001：63-66.

[15] Estabrooks A，Jo T，Japkowicz N. A multiple resampling method for learning from imbalanced data sets [J]. Computational Intelligence，2004，20(1)：18-36.

[16] Elkan C. The foundations of cost-sensitive learning [C]//Proceedings of International Joint Conference on Artificial Intelligence. LAWRENCE ERLBAUM ASSOCIATES LTD，2001，17(1)：973-978.

[17] Zhou Zhihua，Liu Xuying. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. Knowledge and Data Engineering，IEEE Transactions on，2006，18(1)：63-77.

[18] McCarthy K，Zabar B，Weiss G. Does cost-sensitive learning beat sampling for classifying rare classes? [C]//Proceedings of the 1 st international workshop on Utility-based data mining. 2005，21(21)：69-77.

[19] Liu Xuying，Zhou Zhihua. The influence of class imbalance on cost-sensitive learning：An empirical study [C]//Proceedings of Sixth International Conference on. IEEE，2006：970-974.

[20] Mease D，Wyner A J，Buja A. Boosted classification trees and class probability/quantile estimation[J]. The Journal of Machine Learning Research，2007，8：409-439.

[21] Drummond C，Holte R C. C4. 5，class imbalance，and cost sensitivity：Why under-sampling beats over-sampling[C]//Proceedings of Workshop on Learning from Imbalanced Datasets II. 2003.

[22] Batista G E，Prati R C，Monard M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter，2004，6(1)：20-29.

[23] Mani I. knn approach to unbalanced data distributions：A case study involving information extraction [C]//Proceedings of Workshop on Learning from Imbalanced Datasets. 2003.

[24] Hart P E. The Condensed Nearest Neighbor Rule [J]. IEEE Transactions on Information Theory，1968，14：515-516.

[25] Chawla N V，Bowyer K W，Hall L O，et al. SMOTE：synthetic minority over-sampling technique [J]. arXiv preprint arXiv：1106.1813，2011.

[26] Han Hui，Wang Wenyuan，Mao Binghuan. Borderline-SMOTE：A new over-sampling method in imbalanced data sets learning[J]. Advances in Intelligent Computing，2005：878-887.

[27] He Haibo，Bai Yang，Garcia E A，et al. ADASYN：Adaptive synthetic sampling approach for imbalanced learning[C]//Proceedings of IEEE International Joint Conference on IEEE，2008：1322-1328.

[28] Tomek I. Two modifications of CNN [J]. IEEE Trans. Syst. Man Cybern.，1976，6：769-772.

[29] Kubat M，Matwin S. Addressing the curse of imbalanced training sets：one-sided selection[C]//Proceedings of Machine Learning-International Workshop

Then Conference-. Morgan Kaufmann Publishers, Inc., 1997: 179-186.

吴悦昕(1989—),硕士,主要研究领域为数据挖掘和机器学习。
E-mail: wuyuexin@gmail.com

赵鑫(1985—),博士,主要研究领域为网络数据挖掘和自然语言处理。
E-mail: batmanfly@gmail.com

过岩巍(1989—),硕士,主要研究领域为搜索引擎和网络数据挖掘。
E-mail: pkuguoyw@gmail.com

全国知识图谱与语义计算大会(CCKS 2016)在北京隆重召开

2016 年 9 月 19—22 日,全国知识图谱与语义计算大会(CCKS 2016)在北京西郊宾馆隆重召开。本次会议由中国中文信息学会语言与知识计算专业委员会主办。大会分为讲习班和主会两个主要部分,本次讲习班暨中国中文信息学会《前沿技术讲习班》ATT 第三期的主题是知识图谱专题。本次大会吸引了来自全国学术界、产业界从事知识图谱相关研究的 400 多人参加,会议探讨了知识图谱领域的新发现、新技术和新应用,旨在向社会公众介绍知识图谱相关领域的发展趋势和创新成果,进一步推动我国知识图谱技术领域的发展。CCKS 2016 会议的主题是:语义、知识与链接大数据。

会议包括学术讲习班、特邀报告、工业界论坛、评测与竞赛、学术论文、海报及演示等主要环节。其中,前沿技术讲习班邀请了八位国内外知名研究者,分别是:奇点机智的林德康博士、文因互联的鲍捷博士、阿伯丁大学的 Jeff Z. Pan 教授、华东理工大学的阮彤教授、Facebook 的王海勋博士、微软的王仲远博士、南京大学的胡伟博士和南京大学的程龚博士,为大家分享了四个前沿技术讲座。特邀报告邀请了四位国内外的知名研究者,他们分别是:牛津大学的 Ian Horrocks 教授、马普研究所的 Gerhard Weikum 教授、北京理工大学的黄河燕教授和 Facebook 的王海勋博士。工业界论坛邀请了产业界的八位研发人员,分享了实战经验,他们分别是:Franz. 的科学家 Sheng-Yhuan Wu、拓尔思的副总裁刘瑞宝、云知声的 AI 技术专家刘升平、小 i 机器人的陈培华研究员、海云数据的 CTO 赵丹、海翼知的 CEO 丁军、富士通的研究员 Nobuyuki Igata 和图灵机器人的技术负责人韦克礼。

9 月 19—20 日是中国中文信息学会《前沿技术讲习班》ATT 第三期:知识图谱专题。讲习班由清华大学朱小燕教授主持开班仪式。19 号上午第一个讲座是林德康博士和鲍捷博士的《实战中的知识图谱》。19 下午由 Jeff Z. Pan 教授和阮彤教授作讲习班的第二个讲座《Testing and Assessing the Quality of Knowledge Graph》。20 日上午由王海勋博士和王仲远博士作讲习班的第三个讲座《Understanding Short Texts》。20 日下午由胡伟博士和程龚博士作讲习班的第四个讲座《知识图谱的摘要和集成》。

9 月 21—22 日是本次大会的主会。开幕式由专委会主任清华大学李涓子教授主持,首先由中国中文信息学会理事长李生教授致辞,接着由大会主席中国科学院软件研究所孙乐研究员致辞,最后程序委员会主席浙江大学陈华钧教授介绍会议组织情况。

本次会议共邀请了四位海内外知名学者做特邀报告。来自牛津大学的 Ian Horrocks 教授作《Using Semantic Technology to Tackle Industry's Data VarietyChallenge》的报告。德国马普研究所的 Gerhard Weikum 教授作《What Computers Should Know》的报告。北京理工大学的黄河燕教授作《面向基础教育的大数据类人智能答题系统总体设想及其困难与挑战》的报告。Facebook 的王海勋研究员作《Short Text Understanding》的报告。

会议同时设置了学术论文、海报及演示、知识图谱竞赛等环节。知识图谱竞赛部分还邀请了清华大学刘知远博士作《知识表示学习与知识获取》的报告。刘知远博士主要介绍和总结了他们在知识表示和知识获取方面的最新研究进展。

本次大会关注国内外知识图谱研究领域的最新进展,以及工业界的最新技术,对本领域面临的种种挑战性科学问题和关键技术难题展开了深入研讨,为所有与会者带来了一场学术与技术的饕餮盛宴。经语言与知识计算专委会 2016 年工作会议决定,2017 年全国知识图谱与语义计算大会将在四川成都西华大学举办。