情境大数据建模及其在用户行为预测中的应用

吴书，刘强，王亮
中国科学院自动化研究所智能感知与计算研究中心，北京 100190

摘要
随着大数据时代的到来，信息系统收集了海量情境信息，如舆情信息、环境信息、经济信息等。这些情景大数据提供丰富的细节信息，更细致地刻画行为背景以辅助用户行为建模。阐述了两种使用表达学习策略建模一般化情境信息的框架，并针对情境大数据中最常见的时序情境建模问题，使用循环神经网络建模时序情境中的序列依赖关系。

关键词
情境大数据；情境建模；用户建模；行为预测

中图分类号：TP391.4 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016071

Modeling contextual big data for user behavior prediction

WU Shu, LIU Qiang, WANG Liang
Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract
In the big data era, information system has to handle a mass of data of contextual information, such as public opinion, environment information and economic status. Embedded with abundant details of user behavior, contextual information plays a significant role in effectively shaping user character and elaborately modeling user behavior. Two frameworks to model general context information through representation learning and a recurrent model for sequential context scenarios were involved.

Key words
contextual big data, context modeling, user modeling, behavior prediction

1 引言

随着信息技术的快速发展，人类社会进入了全面的信息化时代。伴随着大量网络应用的出现，人们的生活方式发生了改变，越来越多的时间被投入在信息平台上，如个人电脑、智能手机、平板电脑、智能电视等，同时人们的行为方式和习惯也很大程度上被传感器、智能监控等设备收集。随着平台系统收集信息的能力不断增强，大数据时代正在到来。信息系统中收集了用户主动或者被动留下的大量行为数据，同时也收集了大量与用户行为相关的海量情境信息，如社交媒体上的舆情信息、自然环境信息（天气、空气、温度等）、生产经济信息（GDP、生产价格指数、CPI、证券）等数据。在大数据时代的用户分析应用中，越来越多的情景信息能够提供丰富的用户行为细节，更细致更全面地刻画行为发生的背景，有效地辅助用户行为建模。从另一个角度来看，大规模情境建模是一种处理大数据的趋势，它将关联的大数据直接转换为特定目标任务所处环境的复杂情境信息，其作用也越来越重要。

在大数据场景下，当传统行为数据收集极为充分之后，进一步收集行为数据在当前的模型框架下可能无法带来预测性能的大幅度提升。因为当前模型建模的假设大多是针对用户和对象本身，而忽略外在情境因素对用户和对象的影响，更多的用户行为数据也不能拟合出更好的模型参数进而得到更好的算法效果。此时，引入丰富的情境大数据，进一步揭示行为发生的机制则更为重要。目前数据分析领域已经开始重视情境建模，越来越广泛的研究领域在具体任务建模上引入情境大数据，大幅度提升了预测任务的性能。谷歌趋势（Google Trend）将搜索引擎的检索数据引入流感传播的建模过程中[1]。它曾经构建了一套流感预测的系统，通过搜索引擎的检索数据来预估各个时间点流感的状况，在存在外部突发事件时，这套系统的预测结果会因为外部某个事件的刺激而远远偏离真实。后期回到利用疾控中心数据进行预测上，将外部的用户检索数据作为情境信息，获得了更加准确稳定的结果。金融领域也利用经济和社会舆论等情境大数据来辅助建模股价、债券走势的预测，例如美国斯坦福大学和谷歌研究人员训练了一个长短期记忆网络（long-short term memory network，LSTM）模型来预测标准普尔500指数的走势[2]。该模型结合了反映公众情绪和宏观经济的谷歌趋势情境大数据，包含经济类关键词检索结果，获得了远超传统模型的预测效果。

在信息检索和数据挖掘领域的用户行为预测场景中，情境大数据也非常丰富，其中用户行为常常随着这些情境信息的变化而发生改变。例如，当一个用户与小孩在一起时，他可能会倾向于看动画片；当与爱人在一起时，他可能会倾向于看浪漫电影。将情境大数据因素纳入模型构建，能够细致地刻画出用户行为的场景，间接反映出产生用户行为的原因，显著提升行为预测的效果。目前，研究工作主要针对特定的情景信息进行建模，并应用到特定的任务中。例如分析社交媒体上的用户行为，参考文献[3]提出一种结合当前情境下舆情的主题模型，主要运用到与用户兴趣主题相关的领域。

本文主要从两个角度描述情境大数据的建模及其在用户行为预测中的应用。首先，阐述了两种使用表达学习策略建模一般化情境信息的框架，介绍了情境操作张量建模策略[4,5]，同时解释如何将分层表达框架[6]应用在一般化的情境建模场景中。

然后，针对情境大数据中最常见、最重要的时序情境建模问题，介绍基于循环神经网络建模的框架，该框架可用到时序情境建模[7]中，也可用在复杂时序行为建模[8]上。

2 基于表达学习的情境建模框架

在情境信息下预测用户行为最常用的是基于矩阵分解的方法，如张量分解（tensor factorization，TF）[9]和因子分解机（factorization machine，FM）[10]，它假设把一种特定的情境信息当作用户对象之外的另一种实体，并将这种情境信息转化为单独的一个维度，与传统方法中用户对象实体的维度一起进行分解。这类方法仅仅建模了实体和情境信息间的相似度，但这种相似度往往不是很合理。比如，一个用户与工作日这个情境要比与周末这个情境的距离近，同时这类方法难以把握实体和情境交互后的共同潜在特性。一些基于多领域关系预测的模型[11]也可以被用来进行情境感知，它们使用转换矩阵将实体潜在向量从一种情境映射至另一种情境环境下。但是这类方法需要为一个特定的情境信息提供一个转换矩阵，在处理情境大数据时会遇到扩展上的困难。

针对传统模型假设不合理和扩展不足的缺陷，笔者认为实体和情境之间的关系可以使用向量来描述，而不再使用单一的值来表达。这种建模方式能够解决传统框架下的假设局限性，同时利用模型的扩展性可对情境大数据进行建模。本节将介绍两类最新的基于表达学习的情境建模框架：第一个框架通过建模情境信息对用户对象实体的操作，得到实体在当前情境下的表达；第二个框架构建实体和情境信息的层次表达，将它们的交互建模到统一模型中。

2.1 情境信息的表达

传统神经网络语言模型将词表达为连续的语义向量，称之为词嵌入。类似地，也将情境信息转换为向量来表达。同时真实场景中有大量不同类型值的情境信息，如类属型、类属集型和数值型，笔者为它们设计了相应的转换策略。如类属型的情境信息，为每一个特定的情境值学习一个表达；对于类属集型的情境信息，计算出所有元素的平均值作为其表达；对于数值型的情境信息，就为这个情境学习一个表达，任意一个对应的情境值都可以通过乘积操作而得到。有了这3种类型情境信息的向量表达，很多其他类型的情境信息都可以转换为它们的一种，从而得到最终的表达。当用户项目交互中，不同类型值的情境值都被转换为连续值的情境向量之后，需要将交互中的一类情境向量使用加权的方式计算为单一向量，这种向量描述的是当前交互环境中某一类情境信息整体的表达。

2.2 情境操作张量建模框架

受自然语言领域研究的启发，提出一种情境建模方法，称之为情境操作张量（contextual operation tensor，COT）[4, 5]，情景操作张量建模框架如**图1**所示。在自然语言处理的语义分析研究中，名词语义常常被表达为向量，形容词被描述为名词上的操作语义，由操作矩阵来表达这种属性。比如“优质产品”中的名词“产品”被表达为潜在向量，形容词“优质”被表达为矩阵，“优质产品”的联合表达就是矩阵和向量相乘得到的向量表达。假设在用户行为预测中的情境信息具有类似形容词的这种操作属性，能够操作实体的潜在属性，使得情境下的实体新属性不仅能体现出其

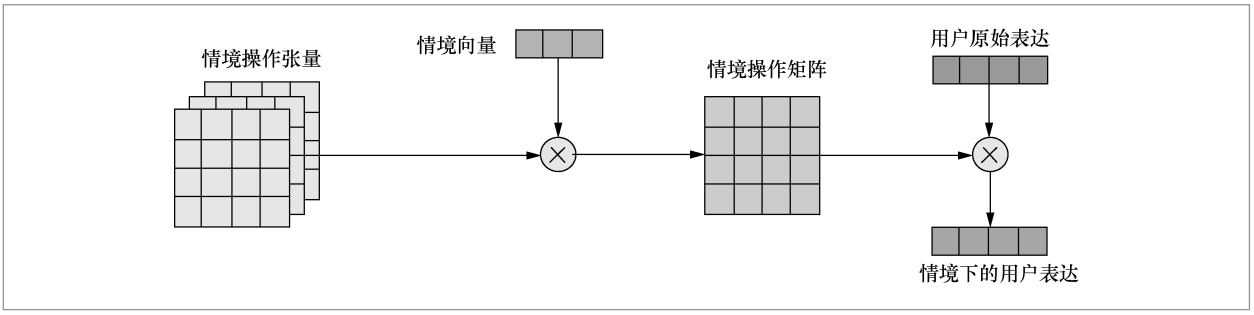


图 1 情境操作张量建模框架

原始属性，也能反映出在特定情境下实体表达上的改变。比如一个用户因为和小孩在一起，这个陪伴的情境信息就改变了用户当下的属性，使其乐意去看动画片。

不同于传统模型中用户和对象都有其固定的不随着情境信息而改变的潜在向量表达，为了描述这种随情境变化的用户和对象的潜在属性，为用户和对象提供了特定情境下的潜在表达。同时将情境信息的潜在语义描述为操作矩阵，它说明对应情境信息有着改变用户对象等实体潜在属性的能力。

因为不同的情境信息常常具有类似的语义，即在对实体属性操作上非常类似，比如人们周末或者在家都会想看小说而非专业书籍。因此通过多个基本的操作矩阵生成情境操作矩阵，这些基本的操作矩阵称为情境操作张量，它们描述的是一些共同的情境语义操作。每个特定情境下的操作矩阵，都可以由它们而产生。因为使用了共有的情境操作张量，这种方式能够有效地减少模型需要拟合参数的数量。

2.3 分层表达情境建模框架

在获取实体和情境表达后，除了将情境信息转换为情境操作矩阵直接作用在实体表达上，也在探索是否有更具扩展性的方式，建模更广泛的情境信息。因此，提出了分层交互表达（hierarchical interaction representation，HIR）模型[6]，将实体之间或者实体和情境之间的交互建模成一个共同的表达，使用一种分层交互表达来描述这种交互，如**图2**所示。

当每种实体和情境信息都使用向量进行表达时，除了待交互的实体和情境表达之外，使用一个三阶张量来获取它们之间的高阶交互。HIR构建了交互的向量表达，

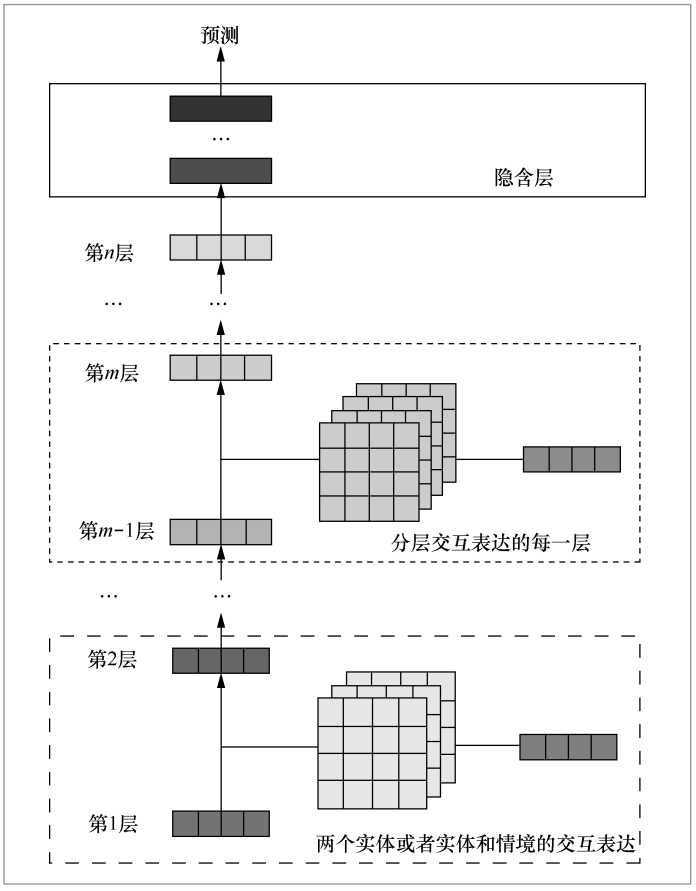


图 2 分层表达情境建模框架

利用张量乘法生成两个实体或者实体和情境的共同表达，然后将这个过程迭代进行，以得到所有实体和情境的最终分层交互表达。HIR具有很好的扩展性，在获得了两个实体或者实体和情境的联合表达之后，可以在框架下建模更多实体和情境的交互。这样的循环操作可以获取所有实体和情境交互作用下的最终表达，这种层次化构建方式得到的交互表达，称之为层次化交互表达。

在获得了最终的层次交互表达之后，可以在其后增加多个隐含层，以挖掘交互的深层隐藏特性，从而进一步增强实体和情境层次交互的表达能力。不同应用中的不同任务都可以基于隐含层的最终表达而构建，根据应用场景运用多种机器学习方法实现预测。在普通推荐、上下文推荐、协同检索、广告点击预测等多个场景进行了实验，实验效果均超过了之前相关领域的最好方法。

3 基于循环神经网络的时序情境建模

情境大数据中的时序情境信息描述的是事件发生最基本的因素，是实际应用系统中建模用户行为的基础，也是最广泛存在的情境信息。较之特定领域的情境信息，时序情境建模更为根本和重要。这类时序情境建模方法具有一般性，可以被引入其他包含时序情境的特定应用领域，例如预测用户签到数据，也可以预测交通堵塞或恐怖组织的攻击行为等。本节将针对时序情境建模展开介绍。

传统的时序情境建模问题受到了广泛的关注，很多研究者开展了一系列研究，相关方法主要包括因子分解方法[9]和基于马尔科夫链[12]的方法。张量因子分解模型将时间当作实体外新的维度，并通过分解得到用户、对象和时间箱体等潜在向量。这类方法在预测那些从来没有或很少出现在训练数据的时间箱体时，会面临冷启动问题。另一方面，基于马尔科夫链的方法已成为最受欢迎的时序预测方法，如个性化因子分解马尔可夫链（factorizing personalized markov chain，FPMC）[12]等。该类基于马尔可夫链的方法都基于马尔可夫假设，只能建模局部序列行为，即相邻行为之间的关系，但序列行为之间常常有着更复杂的关系，需要获取序列高层阶的交互关系，由行为的全局序列特征来做用户行为的预测。

最近循环神经网络（recurrent neural network，RNN）不仅成功应用于自然语言处理领域中的词嵌入(word embedding)[13]，同时也被应用到信息检索领域建模顺序点击预测行为[14]。循环神经网络由输入层、输出层和多个隐藏层组成，其中隐藏层的表示能够动态地随着行为历史而变化，适合用来建模序列信息。然而，该模型只能考虑行为之间的顺序关系，而忽略行为之间的时间间隔信息，这使其在建模具有连续值的时间信息时常遇到困难，而这些具有连续值的时序情境对用户行为的建模往往非常重要。

3.1 时空情境一体化建模

空间和时间描述的是事件的基本因素，即什么时间和什么地点，它们是实际应用中建模用户行为的基础。这些具有连续值的空间和时间情境，对于揭示用户当下的属性有决定性作用，在行为建模上的作用非常重要。因为空间信息的属性非常类似时间信息，将在同一个框架下为它们建模。构建基于RNN的方法建模具有连续值的时空序列信息，称之为时空循环神经网络（spatial temporal recurrent neural

network，ST-RNN)[7]。

时空一体化建模框架如图3所示。传统RNN中每层只考虑一个元素作为输入，ST-RNN将时空序列情境纳入考量，将一个固定时间段内的行为作为一层的输入来建模局部时序信息。同时ST-RNN利用循环结构捕获时序情境信息的周期属性。另一方面，很难给所有的具有连续值的时空信息拟合出对应的转换矩阵，将空间和时间切分为离散的区间。对于某个离散区间中的一个特定时间点，依靠其上界和下界对应的转换矩阵通过线性插值的方式来计算其所对应的转换矩阵，这样ST-RNN就能够使用转换矩阵来表征具有连续值的动态时序信息。类似地，对于一个具有连续值的特定空间信息，也可以通过同样方法生成其转换矩阵。

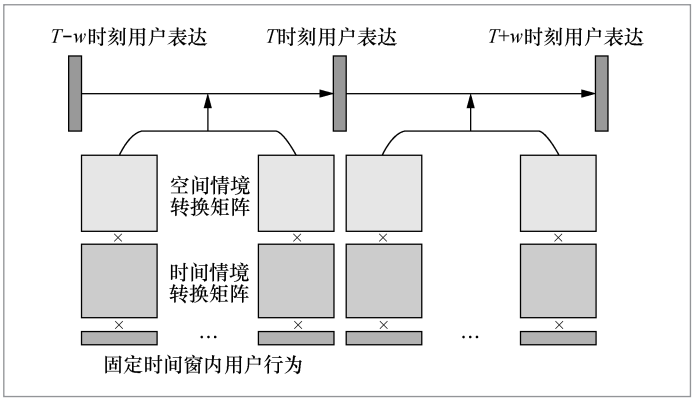


图 3 时空一体化建模框架

3.2 复杂时序情境建模

除了上述的传统时序情境场景外，在现实世界中时序情境往往更为复杂，例如客户常常在同一时刻一次性购买一篮子物品。如何对这种复杂时序情境建模以有效预测用户一篮子购买行为？上述时空情境一体化建模的方法只能实现简单时序场景下用户行为的建模，不能很好地把握单次购买行为中多种物品之间复杂的关联关系。

为了挖掘复杂时序场景中全局序列特征，并揭示用户兴趣的动态变化，依然将循环神经网络作为建模框架引入这项工作[8]。虽然循环神经网络的结构可以捕获所有篮子上用户的全局时序特征，但为了有效建模每次行为内部的复杂情境，将卷积神经网络中的池化操作用于建模篮子本身，提出了动态循环神经网络篮子模型。它的输入实例是由一个特定用户的交易行为组成，每次交易行为由多个对象组成。引入的卷积神经网络中的池化操作能用来获取这些对象整体的表达，能提取出复杂行为对象包含的关键特征信息。笔者使用了最大池化和平均池化两种操作，分别提取所有对象在对应维度上最大值和平均值作为对象整体表达的维度值，复杂情境建模框架如图4所示。

在获得了对象整体表达之后，它将作为输入被放进循环神经网络结构中，然后和输入矩阵进行操作，并与用户之前的隐含状态一起得到下一个状态的用户表达。每个用户的动态表示描述用户属性随着时间推移和与不同篮子进行交互之后潜在属性的变化。池化操作能获得复杂时序行为上最重要的语义属性，同时循环神经网络结构可以从所有用户整体历史交易数据上，获得用户全局序列行为特征。这个框架能取得比传统RNN和基于马尔可夫方法更好的实验效果。

4 结束语

本文介绍了使用表达学习策略建模一般化的情境信息，情境操作张量模型将情境信息看作操作语义，能改变实体在当下情境下的向量表达。分层表达模型使用层

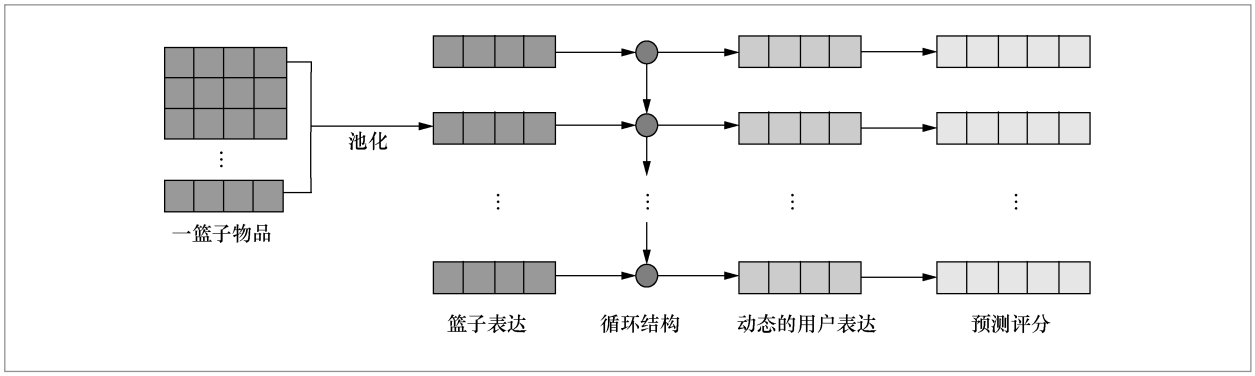


图 4 复杂情境建模框架

次模型来建模实体和情境的交互，获得联合表达。然后，针对最常见的时序情境，介绍如何使用循环神经网络来建模这类信息，并获得当前时序情境建模最好的实验效果。

参考文献:

[1] LAZER D, KENNEDY R, KING G, et al. The parable of Google flu: traps in big data analysis[J]. Science, 2014, 343(6176): 1203-1205.

[2] XIONG R, NICHOLAS E P, SHEN Y. Deep learning stock volatilities with google domestic trends[J]. 2015: arXiv: 1512.04916.

[3] YIN H, CUI B, CHEN L, et al. A temporal context-aware model for user behavior modeling in social media systems[C]// The 2014 ACM SIGMOD International Conference on Management of Data, June 22-27, 2014, Utah, USA. New York: ACM Press, 2014: 1543-1554.

[4] LIU Q, WU S, WANG L. COT: contextual operating tensor for context-aware recommender systems[C]//Twenty-Ninth Conference on Artificial Intelligence, January 25-30, 2015, Austin Texas, USA. [S.l.:s.n.], 2015: 203-209.

[5] WU S, LIU Q, WANG L, et al. Contextual operation for recommender systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(8): 2000-2012.

[6] LIU Q, WU S, WANG L. Collaborative prediction for multi-entity interaction with hierarchical representation[C]//The 24th ACM International on Conference on Information and Knowledge Management, October 18-23, 2015, Melbourne, Australia. New York: ACM Press, 2015: 613-622.

[7] LIU Q, WU S, WANG L, et al. Predicting the next location: a recurrent model with spatial and temporal contexts[C]// Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, USA. [S.l.:s.n.], 2016.

[8] YU F, LIU Q, WU S, et al. A dynamic recurrent model for next basket recommendation[C]//The 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 17-21, Pisa, Italy. New York: ACM Press, 2016: 729-732.

[9] XIONG L, CHEN X, HUANG T K, et al. Temporal collaborative filtering with bayesian probabilistic tensor factorization[C]//The SIAM International Conference on Data Mining, April 29-May 1, Ohio, USA. [S.l.:s.n.], 2010: 211-222.

[10] RENDLE S. Factorization machines with libfm[J]. Acm Transactions on Intelligent Systems and Technology, 2012, 3(3): 57-78.

[11] SINGH A P, GORDON G J. Relational

learning via collective matrix factorization[C]// The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, Las Vegas, USA. New York: ACM Press, 2008: 650-658.

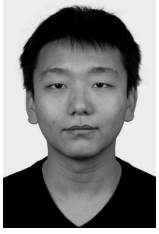
[12] RENDLE S, FREUDENTHALER C, SCHMIDT-THIEME L. Factorizing personalized markov chains for nextbasket recommendation[C]//International Conference on World Wide Web, April 26-30, 2010, Raleigh, USA. New York: ACM Press, 2010: 811-820.

[13] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]//EMNLP, October 25-29, Doha, Qatar. [S.l.:s.n.], 2014(14): 1532-1543.

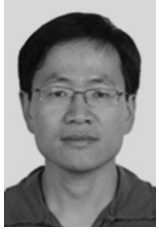
[14] ZHANG Y, DAI H, XU C, et al. Sequential click prediction for sponsored search with recurrent neural networks[J]. Computer Science, 2014: 1369-1375.

作者简介

吴书(1982-),男,中国科学院自动化研究所助理研究员,主要研究方向为数据挖掘和信息检索。先后主持多项国家科研项目,在重要期刊和顶级会议发表论文40余篇。



刘强(1990-),男,中国科学院自动化研究所博士生,主要研究方向为数据挖掘,在顶级会议发表论文多篇。



王亮(1975-),男,中国科学院自动化研究所研究员,博士生导师,IAPR会士和IEEE高级会员,模式识别国家重点实验室副主任,主要研究方向为机器学习、模式识别和计算机视觉。先后主持多项国家科研项目。

收稿日期: 2016-10-28

基金项目: 国家自然科学基金资助项目(No.61403390, No.U1435221)

Foundation Items: The National Natural Science Foundation of China (No.61403390, No.U1435221)