计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700143

基于判断聚合的分布式数据挖掘分类算法研究

李　莉
西南政法大学行政法学院　重庆 401120

摘　要　随着互联网的发展和云计算技术的广泛应用,许多数据存储在不同的服务器上,分布式数据挖掘技术应运而生。智能agent在各自的站点上得到部分挖掘结果,分布式数据挖掘可以将这些部分的挖掘结果聚合成为全局的结果。文中主要处理的是分布式数据挖掘过程中的分类问题,针对一些特征的数据分别存储于不同的数据源上,提出了一种基于判断聚合模型的分类算法。该算法中每一个agent要对一个案例属于某一个目标类的可能性进行判断,然后利用判断聚合模型将这些agent的判断进行聚合,形成全局的分类结果。基于判断聚合模型的分类算法将逻辑和社会选择理论的技术应用于解决分布式数据挖掘的分类问题,这种新的算法不需要大规模地传输和转化数据,节省了传输成本,提高了分类效率,同时有效地保护了数据的安全性。

关键词:分布式数据挖掘;多主体系统;判断聚合模型;逻辑;算法
中图法分类号　TP18

Classification Algorithm of Distributed Data Mining Based on Judgment Aggregation

LI Li
School of Administrative law,Southwest University of Political Science and Law,Chongqing 401120,China

Abstract　With the development of Internet and the wide application of cloud computing,many data sets are stored on different servers,and the distributed data mining comes into being. Each agent gets partial data mining results on its respective site,and distributed data mining could aggregate this part of mining results into a global decision. This paper is focused on the classification issue in the process of distributed data mining. Aiming at some specific data are stored in difference data source,this paper puts forward a classification algorithm based on the judgment aggregation model. Each agent should give its judgment whether a new case belongs to a certain target class,and then use the judgment aggregation model to aggregate the judgments of these agents to form a global classification. This algorithm combines logic and social choice theory technologies together and applies them to the classification problem in distributed data mining. It doesn't need to transfer and transform the data on a large scale,thus saving the transmission cost and improving the efficiency of classification. At the same time,it effectively protects the data security.

Keywords　Distributed data mining,Multi-agent system,Judgment aggregation model,Logic,Algorithm

1　引言

随着计算机技术的普及和应用,企业已经存储了大量的数据,这些数据记载了它们工作流程中各个不同时期的信息,有些企业已经不满足于商业智能(Business Intelligence,BI)的一些简单工具,需要建立个性化的数据挖掘模型从已有的数据中挖掘有用的模式或者发现新的知识,为下一步经营和发展提供决策依据。

近年来,随着物联网、移动互联网以及云计算技术的快速发展,企业数据挖掘需求从起初单一地点单一数据仓库的挖掘亟待推广到对所有地区不同数据点的数据进行挖掘来辅助企业的经营决策,而分布式数据挖掘技术(Distributed Data Mining,DDM)可以解决这一问题。分布式数据挖掘主要处理的是分布式的数据集,即从不同数据点上采集到的数据用于数据挖掘工作,数据挖掘的过程是在局部和全局数据库两个层面上进行的。要想得到全局的模式或者知识,首先要对局部数据库进行挖掘,然后将局部数据库挖掘得到的知识或者模式通过一定的聚合手段聚合成为全局的模式或者知识。分布式数据挖掘的应用范围很广,例如疫情控制、生物制药研究、银行恶意欺诈检测以及客户关系管理与分析等,分布式数据挖掘的过程不需要数据共享,可以阻止数据外泄带来的严重后果,有效地保护数据的安全性。

分布式挖掘对数据集的划分有两种方式:1)平行划分,即所有数据源是同构的,也就是每一个数据点包含同样的属性集;2)垂直划分,即所有的数据源是异构的,也就是每一个数据点仅包含属性集的子集所涵盖的信息。数据挖掘中的传统分类算法,对于同构数据而言,尽管分别位于不同的地理位置上,然而这些数据都是来自于对同一属性集的描述,因此,通过对局部数据库进行挖掘,然后运用简单的汇集手段就可以得到全局的数据挖掘结果;而对异构数据而言,局部数据库进

基金项目:国家社科基金项目(18BZX133);西南政法大学校级项目(2016XZQN-20)
This work was supported by the National Social Science Foundation of China (18BZX133) and University Level Project of Southwest University of Political Science and Law (2016XZQN-20).
通信作者:李莉(395770202@qq.com)

行数据挖掘之后,传统的分类算法无法将局部数据挖掘的结果聚合成为全局的结果,需要新的理论和聚合手段去解决这一问题。

判断聚合模型(Judgment Aggregation Model)是社会选择理论中的一种模型,它主要用于解决聚合问题,判断聚合模型基于逻辑框架建构,针对给定议程 $\mathscr{A}=\{x,y,z,\cdots\}$ 和个体集 $a=\{a_1\cdots,a_n\}(n\geqslant 3)$,个体 a_i 根据自己的聚合规则对议程 $\mathscr{A}$ 中的元素进行判断得到个体判断集,模型以个体判断集作为输入,通过判断聚合机制得到一个一致的聚合结果,判断聚合机制通常表现为特定的判断聚合规则。

本文将基于判断聚合模型对分布式数据挖掘中的分类问题进行研究,特别是解决对于异构数据源数据的分类问题。总体说来,本文尝试在以下几个方面有所突破:

(1)本文分析了大数据环境下的分类问题。从已有的研究文献中来看,同构数据源的分类是比较热门的研究领域,而异构数据源的分类问题显得略为薄弱,本文研究力求为异构数据源的分类问题提供新的解决方案。

(2)本文运用社会选择理论中的判断聚合模型解决了异构数据源数据分类问题,创新地将逻辑和社会选择的方法和技术应用于数据挖掘领域,拓展了逻辑和社会选择领域的研究范围,为分类问题提出了新的研究思路。

(3)本文构建了基于判断聚合模型分布式数据挖掘的新的分类算法,并将这种算法运用于居民健康档案的管理和预测系统中,结果显示,这种新的算法快速准确地将异构数据进行了分类,从案例分析中不难看出,本算法对于防控疫情提出了新的监控思路。

本文第 2 节介绍国内外异构数据源分类问题以及社会选择理论中判断聚合模型的研究现状;第 3 节提出基于判断聚合模型分布式数据挖掘的分类算法,其中主要介绍各个不同 agent 参数的设计以及判断聚合模型的形式化定义;第 4 节将该算法应用于居民健康档案管理领域,对突发疫情的控制过程进行模拟,尝试使用新的算法快速找到感染人群;最后总结全文。

2　相关研究

2.1　分布式环境下基于 agent 的分类算法研究

智能体挖掘(agent mining)概念是一个机器学习、多主体系统和数据挖掘的交叉研究领域,最初由悉尼大学的 Cao 和他的团队提出来,用来解决多主体系统、数据挖掘和机器学习组合技术过程中的交互和整合问题[1]。目前已有一些基于 agent 的分布式分类解决方法的研究成果,其中多数是关于同构数据的分类器的,少数提到了异构数据的分类问题。

在分布式环境下异构分类器方面,Modi 等针对异构数据的分类问题提出了两个基于协作学习的分散算法[2],他们解决方法的核心思想是在不同的数据点上各自基于自身的数据运用相同的算法建立模型,并且同时在局部模型中包含异构学习过程,在学习过程中共享的信息有 agent 的分类模型和训练样例的 *id* 等,但不涉及数据的安全性方面。在学习过程的最后,agent 会经过一个简单投票过程来最终决定集体的预测结果。Santana 和他的研究团队也对协作学习进行了研究,提出一个基于 agent 的神经网络系统来解决异构数据的分类问题[3]。尽管所有 agent 都装入了同样的神经网络算法,agent 之间的区别在于它们学习过程中的参数设定,对于一个没有类标号的新对象而言,为了给出该新对象一致同意的分类,协商的过程中主要依据的是多轮对给定新对象所属类别的置信度,最终,置信度最高的 agent 对新对象给出合理的分类。这种方法避免了大规模数据的共享,一定程度上可以有效地保护数据的安全性。JAM 是一种非协作基于 agent 的系统,它是基于 meta-learning 的一种解决分布式环境下分类问题的方法,这种方法主要应用在银行恶意信用欺诈检测方面。JAM 框架提供了一系列分类算法,这些分类算法会内置到学习 agent 中,而 meta-learning 的 agent 用于整合局部模型形成全局模型。因此,JAM 可以解决分布式环境下异构数据的分类问题,聚合局部模型成为全局模型时,主要运用投票或者 meta-learning 技术。

Kargupta 等开发了 BODHI 系统(BesizingknOwledge through Distributed Heterogeneous Induction)来解决分布式环境下异构数据的分类问题,这个系统也是非协作的基于 agent 的,主要针对异质数据的数据挖掘任务,与之前方法的不同之处在于 BODHI 是要在每一个局部站点的信息中找到有意义的信息片段建立起全局模型,用这种方法替代将局部模型聚合成为全局模型[4-5]。BODHI 的主要思想是,任一函数可以在分布式环境下表达为一个适合的基始函数集,因此,数据建模并不是利用常见的决策树或神经网络等算法。相反地,BODHI 利用正交表达方法在基始空间上学习这些模型,然后再将正交表达方法得到的模型转化成为传统的模式。这种方法不需要大规模的数据会话,并且可以确保局部和全局模型的正确性。

尽管很多分布式数据挖掘系统采用的都是多主体系统(Multi-Agent System,MAS)的架构,但对于异构数据的分类问题有一些基于 MAS 的系统的文献可供参考,例如 Tumer 等提出一种方法从多个异构的数据站点推理得到局部的分类器,这种方法利用基于次序统计的技术将不同分类器对每一个可能类的预测结果进行排序,然后运用特定的聚合手段聚合成为一个一致同意的结果。另外一种处理异构数据的分类问题的方法也是基于 agent 的系统,是 Matatov 等提出的,他们利用朴素贝叶斯将不同局部模型对没有类标号的实例的预测结果聚合成为全局的结果,然而,这种方法主要为了解决数据安全性的问题,即"通过分解解决数据挖掘中安全性问(data mining privacy by decompostion)",也就是通过匿名性的分类将不同的属性集存储在不同的站点上,从而确保数据的安全性[6]。

2.2　社会选择理论中的判断聚合模型研究

社会选择理论是研究集体决策过程的理论,即针对聚合问题提出相应的解决方法,通过建立一系列模型分析决策结果,模型以个体的意见(包括投票、偏好序、判断集或者福利等)为输入,聚合之后的集体意见(包括投票、偏好序、判断集或者福利等)为输出[7-8]。现代社会选择理论研究自 Arrow 提出阿罗不可能定理开始就确立了用逻辑的技术手段去刻画集体决策过程,并在严格的逻辑推理基础上得出社会选择规则无法保证集体决策过程中最起码的公平和理性[9]。偏好聚合模型是 Arrow 在证明不可能性定理过程中建构的,在阿罗不可能定理之后,偏好聚合模型被广泛地应用到社会选择理论中去证明其他的不可能定理。判断聚合模型是 21 世纪以

来新的社会选择理论中新的研究工具，它将集体决策问题放入更一般的逻辑框架内，基于逻辑建立的判断聚合模型可以精确地刻画聚合问题的过程，在严格定义判断聚合规则的基础上推导得出集体判断集。

判断聚合的研究发端于法庭审判，Kornhauser 等于 1986 年在法庭审判过程中发现了判断困境（Discursive Dilemma）[10]，判断困境是典型的个体理性无法达成集体理性的例子，继而 List 等在分析判断困境的基础上建立判断聚合模型，该模型建立在命题逻辑上，其对议程、议题、判断集、判断组合以及判断聚合规则都明确地进行了定义，在此基础上刻画了集体决策的过程，证明了满足无限制定义域、集体理性、匿名性以及系统性这 4 个条件下的判断聚合规则是不存在的，这一基于逻辑的判断聚合模型进一步证明了阿罗不可能定理[11]。判断聚合模型提出之后，有很多学者针对各种不同的情况提出一系列判断聚合规则，文献[12-15]主要针对集体判断集不一致的情形提出判断聚合规则，以期化解这种不一致得到一致的集体判断集。在此基础上，Li 等将判断聚合广泛应用于解决推荐系统、市场营销等领域中的问题[16-17]。

3 一种基于判断聚合模型分布式数据挖掘的分类算法

从相关研究中可以看到，实际的应用中从不同的数据站点上收集所需的数据显然会受到网络带宽和服务器存储量的限制，更为严重的可能直接导致数据泄密，数据的安全性将无法得到保障。然而，目前已有的数据挖掘的技术手段多数是针对集中式的且直接可获得的训练集。本文主要研究的是分布式数据挖掘中针对异构数据的分类算法，即数据的属性集被分别存储在不同的数据点上，分布式数据挖掘问题需要分析分布式数据并且在 agent 的知识范围内产生局部模型，每一个 agent 有着自己的分类目标并且在包含已知分类标签的训练集上建立了更准确的局部模型将新的实例进行分类，然后需要将局部模型形成全局模型，从而得到新的实例的准确分类。然而，针对异构数据的分布式数据挖掘，不同的数据站点形成的局部模型都是不完全的，在形成全局模型的过程中，如何在确保数据安全性的前提下将局部 agent 输出的个体判断集经过全局模型聚合成为集体判断集，是本文要解决的主要问题。

本文中要用到两类 agent，一类是局部 agent，这种 agent 主要为了产生局部模型；另外一类是全局 agent，这一类 agent 主要是将局部模型聚合成为全局模型，从而得到对新的实例的准确分类。

3.1 局部 agent

局部 agent 操作的是本地数据库，它的主要任务是对训练集中的每一个实例的属性值和它所属类之间的关系进行学习，进而产生局部模型，该模型会应用到预测新的实例的分类。对于二元分类任务而言，目标分类是我们感兴趣的方面，被标为“P”，而其余的分类标号是我们不感兴趣的，被标为“O”。

在局部 agent 中输入带有类标号实例的训练集，每一个实例用唯一的 id 标识。每一个实例是被属性集 f 的向量所刻画的，属性集 F 由若干属性组成，任一属性 $f\subseteq f(i=1,\cdots,n)$。给定一组局部 agent 的组合 $\mathcal{A}$ 并且假定是垂直划分的数据，每个 agent a_i 知道每个训练集中实例准确分类，但这些实例的分类只是针对属性集 f 的子集 f_i 的，换句话说，所有 agent 知道训练集中每一个实例的 id 和类标号，但是每一个 agent 只知道描述这些实例的部分属性。作为训练过程的输出，每一个 agent 产生了一个局部模型、一系列实例的预测类标号以及每一个训练集中实例的所属类别的可能性百分比，这些指标可以评估模型的准确性。

agent 通常是建立在机器学习算法的基础之上，即 agent 是利用机器学习算法中的函数，因此，局部 agent 依据机器学习的算法建立局部模型。本文主要探索的是，在分布式分类任务中不同的 agent 是依据不同的机器学习算法建立自己的局部模型，全局模型解决如何将多个分类器得到的结果聚合成为一个实例类标号的问题。局部 agent 内部的分类器算法并不是本文的研究重点。

根据以上分析的思路，局部 agent 可以利用 R 语言中的机器学习算法函数，这些函数已经在 R 语言的 e1071 和 RWeKa 包中，本文假设分布式数据挖掘系统中包含 5 种 agent，每一个 agent 利用以下的一种机器学习的算法建立局部模型，这 5 种训练分类器算法如下。

(1)JRip：从分类实例出发能够归纳出一般的规则。其中重要的算法为 IREP 算法和 RIPPER 算法。重复增量修枝（RIPPER）算法生成一条规则，随机地将没有覆盖的实例分成生长集合和修剪集合，规定规则集合中的每个规则是有两个规则来生成——替代规则和修订规则。

(2)J48：其核心是 ID3。J48 算法保留了 ID3 算法的优点，并对 ID3 算法的缺点进行了改进：1)用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足；2)在树构造过程中进行剪枝；3)能够完成对连续属性的离散化处理；4)能够对不完整数据进行处理。

(3)KNN：如果一个样本在特征空间中的 k 个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。

(4)NB：即朴素贝叶斯分类方法，它是基于贝叶斯定理与特征条件独立假设的分类方法。

(5)SVM：即支持向量机分类算法，一种对线性和非线性数据进行分类的方法。SVM 的原理是：使用一种非线性映射，把原训练数据映射到较高的维上新的维上搜索最佳分离超平面，使用足够高维上的、合适的非线性映射，使两个类的数据总可以被超平面分开。

上述的分类器中只有部分分类器被用作测试，我们认为任一在 R 语言包中给出的分类器都可以被基于 agent 的系统采用，并且由于我们并不关注优化每一种分类器的性能，也不着重比较不同分类器的效率和预测能力，所以对于上述的分类器我们只采用它们的标准参数。

3.2 全局 agent

分布式数据挖掘主要解决的核心问题是如何将各个局部 agent 学到的知识聚合成为全局 agent 的知识。这个问题在异构数据的情况下会变得更加复杂。给定分布式数据点上的局部 agent，每一个局部 agent 在它们的数据点上收集部分属性空间的数据，由于分类任务与属性空间 f 的质量有非常密切的关系，所以，没有局部 agent 可以高质量地学习分类规则得到分类模型，进而将这些局部分类模型聚合成为全局分类

模型会导致分类任务的准确性降低。简单投票规则和元学习的方法在一定程度上无法在聚合过程中体现出所有局部agent的预测结果,在一定程度上忽略了某些局部分类模型得到的结果,使得全局agent在属性空间上信息缺失,直接影响分类结果的准确性和召回率。

本文提出一种新的全局模型聚合机制,基于判断聚合模型的聚合规则可以在一定程度上避免简单多数规则带来的信息缺失问题,因为社会选择理论中的判断聚合模型是在考虑L-一致性和完全性的基础上将个体的判断集聚合成为一个L-一致的集体判断集。本文采用判断聚合模型上的判断聚合规则,相比简单多数规则而言,判断聚合规则可以得到更准确的全局分类结果,但计算的过程会更复杂,因而得出最终分类结果的时间会更长,相较而言效率会降低。选择基于判断聚合机制改进分布式数据挖掘全局模型,需要权衡通讯时间和准确性。

判断聚合模型是以个体判断集为输入、集体判断集为输出,本文中个体判断集可以定义为局部模型产生对某一实例分类的预测,即预测的类标号或者是可能性百分比。局部模型提供了判断聚合模型的输入数据,局部agent将数据传送给全局agent,而判断聚合模型是本文中的全局模型,将局部模型输入的预测类标号或者可能性百分比聚合成为全局的类标号或者可能性百分比。下面定义全局模型,即判断聚合模型。

3.2.1　判断聚合模型

用L来表示命题逻辑,它的语言$\mathscr{L}(L)$具有一般命题逻辑的语义,其语法定义如下:

$\phi::=p\in At\mid\neg\phi\mid\phi\vee\psi\mid\phi\wedge\psi\mid\phi\rightarrow\psi\mid\phi\leftrightarrow\psi\mid\ \bot$

其中,命题联接词有¬(否定)、∨(析取)、∧(合取)、→(蕴涵)以及↔(当且仅当),p表示原子命题,At是原子命题集,根据形成规则由命题连接词和原子命题形成了复合命题,例如$\phi\wedge\psi$,$\phi\vee\psi$和$\phi\rightarrow\psi$等都是复合命题,$\bot$是矛盾式(contradiction),它的否定式是重言式(tautology)。

判断聚合模型一个基本的假设是:所有agent作出的判断既是L-一致的又是完全的。所谓一致的是指作出的判断不能具有矛盾,所谓完全的是指对于所有待断定的命题都要作出判断。相关概念定义如下:

议题(Issue):指等待agent判断的命题。显然,议题可以是$\mathscr{L}$(L)中除了(外的任一命题。议题既不可以是重言式,也不可以是矛盾式,因为它们没有判断的必要。

议程(Agenda):由agent判断的议题以及议题的否定形式构成的集合,记作$\mathscr{A}$。令ϕ是任一议题,则$\phi\in\mathscr{A}$当且仅当$\neg\phi\in\mathscr{A}$,显然,$\mathscr{A}$是非空的且$\mathscr{A}\subseteq\mathscr{L}(L)$,双重否定等值于原命题,即$\neg\neg\phi\equiv\phi$。对议程进行这样的规定是为了保证判断聚合中所要断定的命题都在议程中且是可以被语言$\mathscr{L}(L)$所刻画。令$[\mathscr{A}]\subseteq\mathscr{A}$,但$[A]$中只包含议题本身,即$A=[A]\cup\{\neg\phi\in[\mathscr{A}]\}$,这样的$[\mathscr{A}]$称为预议程,设定$[\mathscr{A}]$是为了方便后面的计算和说明。

一致性(Consistent):如果对于任一公式集$S\neg\mathscr{L}(L)$,如果不存在一个议题ϕ使得$S\vdash_L\phi\wedge\neg\varphi$,则称$S$具有一致性或者说$S$是L-一致的。

完全性(Completeness):对于任一公式集$S\subseteq\mathscr{L}(L)$,任一集合$S\subseteq\mathscr{A}$,如果对于$\mathscr{A}$中的议题$\varphi\in\mathscr{A}$,$\varphi\in S$或者$\neg\varphi\in S$,则称S具有完全性或者说S是完全的。

判断集(Judgment set):判断集是与局部agent与全局agent相关的概念,用J表示。令$\mathscr{N}$表示所有的局部agent,$\mathscr{N}=\{a_1,a_2,\cdots,a_n\}(n\geqslant2)$,则任一局部agent $a_i(1\leqslant i\leqslant n)$的判断集$J_{ai}$是一个一致的且完全的集合,且$J_{ai}\subset\mathscr{A}$。判断集的一致性保证了局部或全局agent所作出的判断不产生矛盾,判断集的完全性保证了个体对议程中每一项议题都作出判断。

判断组合(Judgment profile)用$\mathscr{P}$表示,是由所有局部agent的判断集组成的n元组,可表示为$\mathscr{P}=(J_1,\cdots,J_n)\in\mathbb{D}^n(\mathscr{A})$。我们用$N(\mathscr{P},\phi)$表示判断组合$\mathscr{P}$中有多少个局部agent的判断集中包含$\phi$,也就是$N(\mathscr{P},\phi)=\#\{i\mid J_i\in\mathscr{P},\phi\in J_i\}$,也称作议题$\phi$的权重。

3.2.2　判断聚合规则

定义1(判断聚合规则,Judgment aggregation rules)　令$\mathscr{N}$是所有的局部agent,$\mathscr{N}=\{a_1,a_2,\cdots,a_n\}(n\geqslant2)$,议程$\mathscr{A}\subseteq\mathscr{L}(L)$,$J(\mathscr{A},L)$表示议程$\mathscr{A}$上所有判断集的集合,$J(\mathscr{A},L)^n$表示所有个体的判断集构成的$n$元判断组合$\mathscr{P}=(J_{a_1},\cdots,J_{a_n})$$(J_{a_i}\in\mathbb{D}(\mathscr{A}))$。一个判断聚合规则就是一个以$J(\mathscr{A},L)^n$为定义域、以$J(\mathscr{A},L)$为值域的函数:

$F:J(\mathscr{A},L)^n\rightarrow J(\mathscr{A},L)$

定义2(多数聚合规则,Majoritartian aggregation rule)　这一规则输出的判断集是由多数局部agent支持的议题构成,即对任一$J\in J(\mathscr{A},L)$,$J=\{\phi\in\mathscr{A}\mid N(\mathscr{P},\phi)>1/2\}$,记作$F_{maj}$。多数聚合规则的阈值暂定为1/2,根据应用的需求可调高或者调低。

定义3(权重最大子议程判断规则,Maxweight sub-agenda rule)　对于任一议程$\mathscr{A}$,以及议程$\mathscr{A}$上的判断组合$\mathscr{P}=(J_{a_1},\cdots,J_{a_n})\in\mathbb{D}^n(\mathscr{A})$,任一议题$\phi$在判断组合中的权重是$N(\mathscr{P},\phi)=\#\{a_i\mid J_{ai}\in\mathscr{P},\phi\in J_{ai}\}$,判断集$J$在判断组合$\mathscr{P}$上的权重为$\mathrm{MWA}(P)=\arg\max_{J\in D(\mathscr{A},\Gamma)}W_{\mathscr{P}}(J)$。

3.3　算法描述

假设分布式系统由n个局部agent构成,$\mathscr{N}=\{a_1,\cdots,a_n\}$$(n\geqslant2)$和一个全局agent组成,本文要解决的是新实例集的分类问题,即对于一个新实例集$M=\{R_1,\cdots,R_n\}(n\geqslant2)$,实例集中的实例被属性空间中的所有属性所刻画,由于各个属性都被分布在不同的数据站点上,因而新实例集M根据垂直划分的标准将不同属性的数据送入不同的局部数据库中。由于之前局部agent在对训练集的分类过程中已经建立起局部模型,现在局部模型要对新实例集中的各条数据预测其分类,目标分类设置为"Positive"(记作P),其余分类设置为"Others"(记作O),每一个局部模型要对$M=\{R_1,\cdots,R_n\}(n\geqslant2)$中的所有新实例$R_i$是否属于目标分类进行判断,目标类标号和支持度百分比一起构成了局部agent的预测结果。根据判断聚合模型的定义,首先确定议程$\mathscr{A}=\{S_{R_{i,j}}\mid R_i\in M,j=P\}\cup\{\neg S_{R_{i,j}}\mid R_i\in M,j=O\}$,我们用$S_{R_{i,j}}$表示任一局部agent a_i对任一新实例R_i是否属于某一分类的判断,$S_{R_{i,j}}\in\mathscr{A}$,此时是$S_{R_{i,j}}$议题。所有局部agent给出的预测结果根据判断聚合模型就是每一个局部agent对新实例的分类的个体判断集J_{a_i}。然后,局部agent将个体判断集J_{a_i}传送给全局agent,由全局agent将局部agent的个体判断集J_{a_i}进行聚合,运用一定的聚合规则得到全局agent对任一R_i的分类,最终得到是新实例集M中新实例的分类,即将输入的不同数据源的数据运用

判断聚合模型进行分类，把所有的信息进行充分整合之后得到两个分类，一个是对决策有用的分类，表示为“真”，一般用“1”表示，将其余分类的都归为一类，表示为“假”，一般用“0”表示。算法流程如图 1 所示。

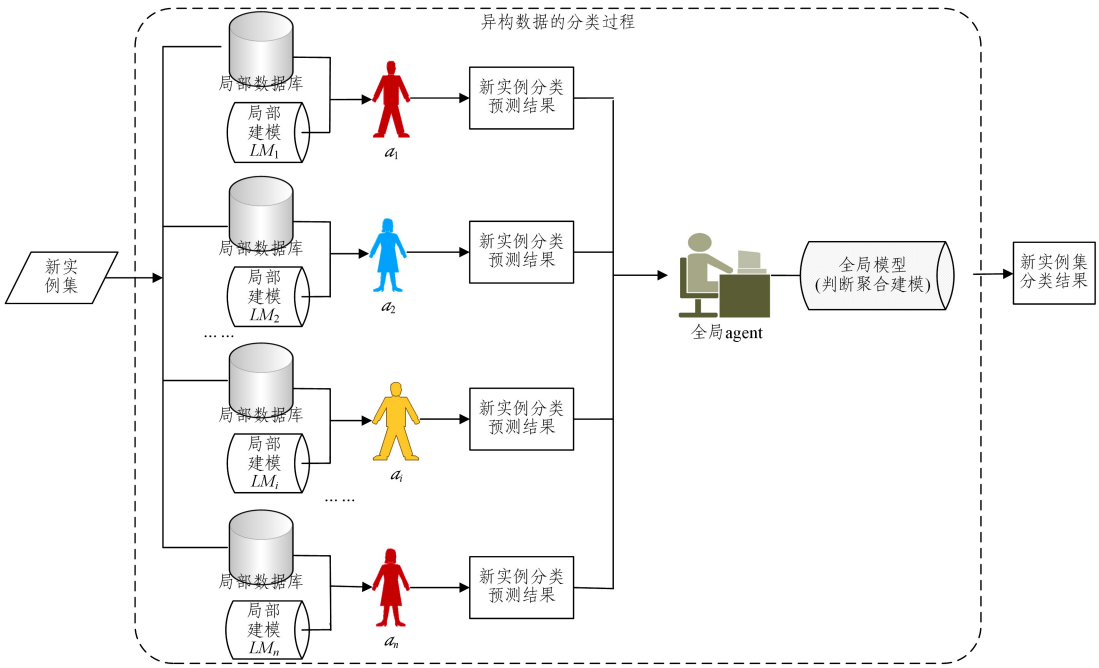

图 1 算法流程图

Fig. 1 Flow chart of proposed algorithm

从图 1 看到，新实例被属性集中所有属性所刻画，而由于数据是异构的，属性集分别来自于不同的数据源，因而数据也存储于不同的局部数据库中，新实例集中的 id 是所有局部数据库的公共关键字，这个 id 有可能是身份证号码、产品条形码或者是唯一确定的一个号码，一般为浮点字段。经过局部模型之后，可以得到每一个 id 标识的新实例属于目标分类的可能性百分比，局部 agent 将这个分类信息传送给全局 agent，全局 agent 接收到分类信息之后作为全局模型的输入，全局模型是上文中定义的判断聚合模型，依据一定的判断聚合规则得到一个唯一的分类结果，输出的分类结果就是新实例集中所有实例的分类信息，这个分类信息可用作进一步决策的依据。

在实际应用过程中，考虑到通讯时间的问题，一般全局 agent 会先设定支持度阈值，如果将支持度阈值设定为 50%，通过多数聚合规则和权重最大子议程判断规则之后，全局 agent 可以得到 L-一致且完全的判断集，也就是对每一个实例的分类有了明确的判断，这一次分类任务就已经完成；如果在这个支持度阈值之下，有部分实例的分类存在争议，此时需要对支持度阈值进行调整，若分类任务涉及疫情控制这种危害性比较大的事件，可以将阈值降低至所需的程度，以期将危害降低到最低，再运用多数聚合规则和权重最大子议程判断规则进行聚合，直至得到 L-一致且完全的判断集，结束分类任务。

4 实例分析

现在我们以某城居民健康档案系统为例来说明本文提出的分类算法，分布式分类系统是以居民的身份证号码为统一的 id，即各个数据点的数据库共享 id 作为所有居民的身份标识，目前共有 4 个数据源，它们分别是：卫生局的居民基本健康信息、医院的居民诊疗信息、药店的居民购药记录以及社区诊所居民体检信息，假设一种名字为 ARS 的呼吸系统传染病入侵该城，通过 ARS 病症的典型症状对该城的所有常住居民进行筛查，预测易感染的居民，并提醒他们到就近的医院进行检查，从而防止传染性疾病的大面积扩散。

卫生局的居民基本健康信息包括的个人基本信息的特征包含有：id、姓名、性别、年龄、证件类型、证件号码、居住地以及手机号码，关于个人基本健康信息的特征有：药物过敏史、外伤史、输血史、疾病史、家族遗传病史、家族病史-父母、家族病史-兄弟姐妹、家族病史-子女以及残疾情况，同时还要加入该次分类任务的两个特征：类标号以及可能性百分比，类标号和可能性百分比会通过 agent a_1 传送给全局 agent 。医院的居民诊疗信息的个体基本信息特征有：id、姓名、性别、年龄、科室；疾病信息特征有：体温、血压、心率以及症状；诊断结果信息特征有：购买日期、药品编码、药品名称、剂量，同时还要加入该次分类任务的两个特征：类标号以及可能性百分比，类标号和可能性百分比会通过 agent a_2 传送给全局 agent。药店居民购药记录信息特征有：id，药品编码、药品名称以及数量，同时还要加入该次分类任务的两个特征：类标号以及可能性百分比，类标号和可能性百分比会通过 agent a_3 传送给全局 agent。社区诊所个体基本信息特征有：id、姓名、性别、年龄、手机号码，结果特征有：心电图、血常规、五官科、血相化验，同时还要加入该次分类任务的两个特征：类标号以及可能性百分比，类标号和可能性百分比会通过 agent a_4 传送给全局 agent。

医生对 ARS 的致病菌的临床研究发现，感染 ARS 的患者会有父母呼吸系统方面的遗传特征，其子女也存在遗传特征；会有高热心率超过 100 的临床症状；并且 ARS 的致病菌潜伏期为 10 天到 20 天，患者在潜伏期有类似感冒的症状，患者会 30 天内在附近的药店购买退热镇痛类药物；如果患者在潜伏期和发病期进行过体检，那么会有明显的白血球升高、高热并且心率超过 100 的症状。从医生的临床研究来看，现在

要对该城的常住人口进行分类,即可能感染 ARS 病菌的是我们目前最为关心的,这一类成为目标分类,记作“P”,不可能感染 ARS 病菌的分类被视作其他分类,记作“O”。

现在对某社区的一栋博士后公寓楼 12 位居民进行排查之后,卫生局、医院、药店和社区诊所 4 个局部 agent 将他们收集到的居民信息放入各自的数据库中,运用上述的机器学习算法进行计算之后,将 *id*、类标号和支持度这 3 个特征的值传送给全局 agent,全局 agent 得到的数据如表 1 所列,其中类标号“P”表示目标分类,“O”表示其他分类,本文为了保护居民的身份证号码信息,所有的 *id* 用 $R_1,\cdots,R_n$ 代替。

表 1　全局 agent 得到的 12 位居民的数据

Table 1　Records of 12 residents obtained from global agent

Id	卫生局(agent1)		医院(agent2)		药店(agent3)		体检中心(agent4)	
	类标号	支持度/%	类标号	支持度/%	类标号	支持度/%	类标号	支持度/%
R_1	P	70	P	56	P	90	O	15
R_2	O	97	P	15	O	80	O	73
R_3	P	93	O	16	O	35	P	5
R_4	P	25	P	89	P	58	O	36
R_5	P	97	O	13	P	89	P	71
R_6	P	83	O	24	O	23	P	21
R_7	P	17	P	56	P	90	O	15
R_8	O	97	P	85	O	80	O	73
R_9	O	94	P	27	P	78	O	63
R_{10}	P	41	P	58	P	76	O	72
R_{11}	O	97	P	13	O	77	O	83
R_{12}	P	99	O	4	O	41	P	95

此次分类任务的目标是从居民中找出那些携带或者感染了 ARS 的患者,也就是在分类过程中我们更关注那些类标号为“P”的人,因为这些人非常有可能已经感染病毒。现在设置感染 ARS 支持度阈值为 50%,即如果类标号为“P”且支持度大于或等于 50%,则我们认为该位居民已经感染 ARS;如果类标号为“P”且支持度小于 50%,则我们认为该位居民并没有感染 ARS;如果类标号为“O”且支持度大于或等于 50%,则我们认为该位居民并未感染 ARS;如果类标号为“O”且支持度小于 50%,则我们认为该位居民已经感染 ARS。

现在以 R_1 为例来说明全局 agent 如何将 4 个数据源传来的对 R_1 的分类进行聚合。从表 1 中可以看出,卫生局端 agent a_1 传送来的信息显示,R_1 感染 ARS 的类标号为“P”,并且支持度是 10%,也即是 R_1 感染 ARS 的可能性是 10%;医院端 agent a_2 传送来的信息显示,R_1 感染 ARS 的类标号是“P”且支持度是 56%,也即是 R_1 感染 ARS 的可能性是 56%;药店端 agent a_3 传送来的信息显示,R_1 感染 ARS 的类标号是“P”且支持度是 90%,也即是 R_1 感染 ARS 的可能性高达 90%;社区诊所端 agent a_4 传送来的信息显示,R_1 是感染 ARS 分类类标号是“O”且支持度是 85%,也即是 R_1 感染 ARS 分类的可能性是 15%。

上例中有 4 个数据源,即卫生局、医院、药店和社区诊所,我们分别用 a_1,a_2,a_3 和 a_4 表示 4 个数据源的局部 agent,局部 agent 的集合 $\mathcal{N}=\{a_1,a_2,a_3,a_4\}$,某社区这栋博士后公寓的所有居民的集合是 $R=\{R_1,\cdots,R_{12}\}$,所有 agent 要对居民集合中的所有居民是否感染 ARS 病菌进行判断,例如,a_1 给出 R_1 的类标号是“O”,a_2 给出 R_1 的类标号是“P”,以此类推,每一个局部 agent 都对每一位居民的类标号进行了预测,因而根据判断聚合模型的定义,议程 $\mathcal{A}=\{S_{R_{i,j}}\mid R_i\in M,j=P\}\bigcup\{\neg S_{R_{i,j}}\mid R_i\in M,j=O\}$,$S_{R_{i,j}}$ 表示任一局部 agent a_i 对任一居民 R_i 是否感染 ARS 的判断,也即是 $S_{R_{i,j}}\in\mathcal{A}$,表示议题,同时系统要求所有局部 agent 需要对这栋楼中的所有居民是否感染 ARS 进行判断。分类标号为“P”表示已经感染,“O”表示未感染。将图 2 中所有数据依据感染阈值设置和判断聚合模型的定义转化为图 3 中的判断组合,我们只取 $S_{R_i,P}$ 本身情况下的议程,根据议程的定义称作预议程[$\mathcal{A}$],即如果任一 agent a_i 判定 R_i 的类标号是 P,那么命题 $S_{R_i,P}$ 为真,如果任一 agent a 判定 R_i 的类标号是 O,那么命题 $S_{R_i,P}$ 为假。表 2 中“1”表示真,“0”表示假。

表 2　判断组合 $\mathcal{P}$

Table 2　Judgment profile $\mathcal{P}$

[$\mathcal{A}$]	$S_{R_1,P}$	$S_{R_2,P}$	$S_{R_3,P}$	$S_{R_4,P}$	$S_{R_5,P}$	$S_{R_6,P}$	$S_{R_7,P}$	$S_{R_8,P}$	$S_{R_9,P}$	$S_{R_{10},P}$	$S_{R_{11},P}$	$S_{R_{12},P}$
a_1	1	0	1	0	1	1	0	0	0	0	0	1
a_2	1	0	1	1	1	1	1	1	0	1	0	1
a_3	1	0	1	1	1	1	1	0	1	1	0	1
a_4	1	0	0	0	1	0	1	0	0	0	0	1

从表 2 的判断组合中可以得到每一个局部 agent 的个体判断集,即 agent a_1 的个体判断集 $J_{a_1}=\{S_{R_1,P},S_{R_2,P},S_{R_3,P},\neg S_{R_4,P},S_{R_5,P},S_{R_6,P},\neg S_{R_7,P},\neg S_{R_8,P},\neg S_{R_9,P},\neg S_{R_{10},P},\neg S_{R_{11},P},S_{R_{12},P}\}$;agent a_2 的个体判断 $J_{a_2}=\{S_{R_1,P},\neg S_{R_2,P},S_{R_3,P},S_{R_4,P},S_{R_5,P},S_{R_6,P},S_{R_7,P},S_{R_8,P},\neg S_{R_9,P},S_{R_{10},P},\neg S_{R_{11},P},S_{R_{12},P}\}$;agent a_3 的个体判断集 $J_{a_3}=\{S_{R_1,P},\neg S_{R_2,P},S_{R_3,P},S_{R_4,P},S_{R_5,P},S_{R_6,P},\neg S_{R_7,P},\neg S_{R_8,P},S_{R_9,P},S_{R_{10},P},\neg S_{R_{11},P},S_{R_{12},P}\}$;agent a_4 的个体判断集为 $J_{a_4}=\{S_{R_1,P},\neg S_{R_2,P},\neg S_{R_3,P},\neg S_{R_4,P},S_{R_5,P},\neg S_{R_6,P},S_{R_7,P},\neg S_{R_8,P},\neg S_{R_9,P},\neg S_{R_{10},P},\neg S_{R_{11},P},S_{R_{12},P}\}$。

根据权重最大子议程判断聚合方法的定义,计算每一项议题局部 agent 的支持数,即每项议题的权重:$N(\mathcal{P},S_{R_1,P})=4$;$N(\mathcal{P},S_{R_1,O})=0$;$N(\mathcal{P},S_{R_2,O})=0$;$N(\mathcal{P},S_{R_2,N})=4$;$N(\mathcal{P},S_{R_3,P})=3$;$N(\mathcal{P},S_{R_3,O})=1$;$N(\mathcal{P},S_{R_4,P})=2$;$N(\mathcal{P},S_{R_4,N})=2$;$N(\mathcal{P},S_{R_5,P})=4$;$N(\mathcal{P},S_{R_5,O})=0$;$N(\mathcal{P},S_{R_6,P})=3$;$N(\mathcal{P},S_{R_6,O})=1$;$N(\mathcal{P},S_{R_7,P})=3$;$N(\mathcal{P},S_{R_7,O})=1$;$N(\mathcal{P},S_{R_8,P})=1$;$N(\mathcal{P},S_{R_8,O})=3$;

$N(\mathscr{P}, S_{R_9,P})=1$；$N(\mathscr{P}, S_{R_9,O})=3$；$N(\mathscr{P}, S_{R_{10},P})=2$；$N(\mathscr{P}, S_{R_{10},O})=2$；$N(\mathscr{P}, S_{R_{10},P})=0$；$N(\mathscr{P}, S_{R_{10},O})=4$；$N(\mathscr{P}, S_{R_{12},P})=4$；$N(\mathscr{P}, S_{R_{12},O})=0$。

局部 agent 对所有议题的权重的计算，我们发现 R_4 和 R_{10} 比较特别，这两位居民从 4 个不同的数据点的采集回来的数据发现，两个数据点返回的数据认为他们可能感染 ARS 病毒，两个数据点返回的数据认为他们并未感染 ARS 病毒，为了有效地控制疫情，对于这种平局情形，我们将感染阈值降低到 40%，即大于或等于 40%我们认为该位居民未感染 ARS，小于 40%认定该位居民未感染 ARS。此时，对于 R_4 而言，社区诊所给出的类标号是“O”且支持度是 36%，由于降低了感染阈值，R_4 没有感染 ARS 的可能性是 36%，即他感染了 ARS 的可能性是 64%，此时社区诊所传给全局 agent 的类标号应该为“P”而不是原来的“O”，相应地，表 2 中社区诊所agent a_4 对 $S_{R_4,P}$ 的判断应为真，如此一来得到 $S_{R_4,P}$ 和 $S_{R_1,N}$ 的权重为：$N(\mathscr{P}, S_{R_4,P})=3$；$N(\mathscr{P}, S_{R_4,N})=1$；同理，$R_{10}$ 也是类似的情况，卫生局给出的类标号是“P”且支持度是 41%，降低感染阈值之后，卫生局传给全局 agent 的类标号应为“P”，因为超过 40%的支持度就被认定为感染 ARS，相应地，表 2 中社区诊所 agent a_1 对 $S_{R_{10},P}$ 的判断应为真，如此一来得到 $S_{R_{10},P}$ 和 $S_{R_{10},O}$ 的权重为：$N(\mathscr{P}, S_{R_{10},P})=3$；$N(\mathscr{P}, S_{R_{10},O})=1$。通过降低感染阈值，我们解决了 R_4 和 R_{10} 的局部 agent 分类标号平局的情形，由此可以得到新的判断组合，如表 3 所列。

表 3 判断组合 $\mathscr{P}'$

Table 3 Judgment profile $\mathscr{P}'$

$[\mathscr{A}]$	$S_{R_1,P}$	$S_{R_2,P}$	$S_{R_3,P}$	$S_{R_4,P}$	$S_{R_5,P}$	$S_{R_6,P}$	$S_{R_7,P}$	$S_{R_8,P}$	$S_{R_9,P}$	$S_{R_{10},P}$	$S_{R_{11},P}$	$S_{R_{12},P}$
a_1	1	0	1	0	1	1	0	0	0	0	0	1
a_2	1	0	1	1	1	1	1	1	0	1	0	1
a_3	1	0	1	1	1	1	1	0	1	1	0	1
a_4	1	0	0	0	1	0	1	0	0	0	0	1

由表 3 新的判断组合可以看到，所有议题的权重变为：$N(\mathscr{P}', S_{R_1,P})=4$；$N(\mathscr{P}', S_{R_1,O})=0$；$N(\mathscr{P}', S_{R_2,P})=0$；$N(\mathscr{P}', S_{R_2,O})=4$；$N(\mathscr{P}', S_{R_3,P})=3$；$N(\mathscr{P}', S_{R_3,O})=1$；$N(\mathscr{P}', S_{R_4,P})=3$；$N(\mathscr{P}', S_{R_4,O})=1$；$N(\mathscr{P}', S_{R_5,P})=4$；$N(\mathscr{P}', S_{R_5,O})=0$；$N(\mathscr{P}', S_{R_6,P})=3$；$N(\mathscr{P}', S_{R_6,O})=1$；$N(\mathscr{P}', S_{R_7,P})=3$；$N(\mathscr{P}', S_{R_7,O})=1$；$N(\mathscr{P}', S_{R_8,P})=1$；$N(\mathscr{P}', S_{R_8,O})=3$；$N(\mathscr{P}', S_{R_9,P})=1$；$N(\mathscr{P}', S_{R_9,O})=3$；$N(\mathscr{P}', S_{R_{10},P})=3$；$N(\mathscr{P}', S_{R_{10},O})=1$；$N(\mathscr{P}', S_{R_{11},P})=0$；$N(\mathscr{P}', S_{R_{11},O})=4$；$N(\mathscr{P}', S_{R_{12},P})=4$；$N(\mathscr{P}', S_{R_{10},O})=0$。根据权重最大子议程判断聚合规则的定义，我们要找到一个总的支持权重最大且 L-一致的集体判断集，由于所有的议题都是原子命题，它们并不存在合取、析取、蕴涵等逻辑关系，因此，可以得到总的权重最大的值是：$W(\mathscr{P}')=(\{S_{R_1,P}, \neg S_{R_2,P}, S_{R_3,P}, S_{R_4,P}, S_{R_5,P}, S_{R_6,P}, S_{R_7,P}, \neg S_{R_8,P}, \neg S_{R_9,P}, S_{R_{10},P}, \neg S_{R_{11},P}, S_{R_{12},P}\})=41$，此时这个集体判断集 $\{S_{R_1,P}, \neg S_{R_2,P}, S_{R_3,P}, S_{R_4,P}, S_{R_5,P}, S_{R_6,P}, S_{R_7,P}, \neg S_{R_8,P}, \neg S_{R_9,P}, S_{R_{10},P}, \neg S_{R_{11},P}, S_{R_{12},P}\}$ 就是全局 agent 对每一位居民感染状况的判断，其中除了 R_2，R_8，R_9 和 R_{11} 外，其余的都是 ARS 病毒的感染者或者是携带者，他们需要进一步的确诊。

在这个案例中，我们不难发现逻辑与社会选择理论在分布式数据挖掘过程中有着独特的作用。首先，判断聚合模型可以将个体判断集聚合成为一个 L-一致的集体判断集，在这个过程中，只需要各个站点传输来各自的预测结果，即各个局部 agent 的个体判断集，通过求得权重最大的 L-一致集体判断集的规则，得到一个 L-一致的集体判断集，这个集体判断集可以作为下一步进行确诊的依据；其次，当出现 4 个数据源的局部 agent 传来的结果是平局的情形，我们可以灵活地调整阈值来得到一致集体判断集，调整阈值的依据通常是事件对社会的危害程度，如果在上例中的疫情控制情况下，我们宁可选一个较低的阈值，找到那些受感染的人；最后，由于目前通常使用的判断聚合模型是基于命题逻辑的，集体判断集的一致性是最基本的要求，进而决策结果绝不会出现矛盾的情况，对疫情管理部门的决策有着重要的意义。

结束语 本文提出了一种基于判断聚合模型的异构数据的分类方法，该方法用判断聚合建模将若干局部 agent 输入的有关实例的分类信息聚合成一个 L-一致的判断集，这个判断集呈现的就是实例的最终分类信息。通过实例分析表明，基于判断聚合的异构数据分类算法比传统的分类算法更加简洁高效，同时由于判断聚合模型是建立在命题逻辑的基础之上，得到的结果必然是有效的，对于实务部门的下一步决策有着切实的指导作用，这充分说明了该算法的优越性。

下一步工作可以从以下几个方面拓展本文的研究。首先，将探索局部 agent 的分类算法，并测试已有算法的准确性，同时改进这些算法中的不足之处。其次，可以尝试将社会选择中更多的模型用于全局 agent 的建模中，例如偏好聚合模型等，并用实验评估判断聚合模型和偏好聚合模型之间的优劣。最后，尝试运用本文中的分类算法到更多的场景中，比如信用卡风险防控、犯罪监控、学生作弊监控以及学术不端检测等方面。

参 考 文 献

[1] CAO L, WEISS G, YU P S. A brief introduction to agent mining [J]. Autonomous Agent and Multi-Agent Systems, 2012, 25(3): 419-424.

[2] MODI P J, SHEN W M. Collaborative multiagent learning for classification tasks[C]// Proceedings of the Fifth International Conference on Autonomous Agent. ACM, 2001: 37-38.

[3] SANTANA L E A, CANUTO A M P, JUNIOR J C X, et al. A comparative analysis of data distribution methods in an agent-based neural system for classification tasks[C]// Sixth International Conference on Hybrid Intelligent Systems (HIS'06). IEEE, 2006: 9-9.

[4] KARGUPTA H, HUANG W, SIVAKUNMAR K, et al. Distributed clustering using collective principal component analysis [J]. Knowledge and Information Systems, 2001, 3(4): 422-448.

[5] PARK B H, KARGUPTA H. Distributed data mining: Algorithms, systems, and applications [M]// Data Mining Handbook. 2002.

(下转第 466 页)

region-based selection in evolutionary multiobjective optimization[C]// Conference on Genetic & Evolutionary Computation. 2001.
[13] ZHANG S, LIU W Q, ZHAO N. Research of Consensus in Multi-agent Systems on Complex Network [J]. Journal of Frontiers of Computer Science, 2019, 46(4): 101-105.
[14] ANGELINI L, BOCCALETTI S, MARINAZZO D, et al. Identification of network modules by optimization of ratio association [J]. Chaos, 2007, 17(2): 175.
[15] DEB K, PRATAP A, AGARWAL S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197.
[16] ZHOU A, QU B Y, LI H, et al. Multiobjective evolutionary algorithms: A survey of the state of the art[J]. Swarm & Evolutionary Computation, 2011, 1(1): 32-49.
[17] LI J Y, ZHOU J G, GUAN J H, et al. Research progress of spectral clustering algorithm [J]. Journal of Intelligent Systems, 2011, 6(5): 405-414.
[18] NEWMAN M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582.
[19] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 19-44.
[20] ZHANG L, PAN H, SU Y, et al. A Mixed Representation-Based Multiobjective Evolutionary Algorithm for Overlapping Community Detection [J]. IEEE Transactions on Cybernetics, 2017, 47(9): 2703-2716.
[21] ZHANG X, ZHOU K, PAN H, et al. A Network Reduction-Based Multiobjective Evolutionary Algorithm for Community Detection in Large-Scale Complex Networks[J]. IEEE Transactions on Cybernetics, 2018, 50(2): 703-716.
[22] LANCICHINETTI A, FORTUNATO S, RADICCHIF. Benchmark graphs for testing community detection algorithms[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2008, 78(4 Pt 2): 046110.
[23] DANON L, DÍAZGUILERA A, DUCH J, et al. Comparing community structure identification[M]// Research and Innovation. Policies and Strategies in the Age of Globalization. 2005.
[24] LUSSEAU D. The emergent properties of a dolphin social network[J]. Proceedings Biological Sciences, 2003, 270(2): S186.
[25] ZACHARY W W. An Information Flow Model for Conflict and Fission in Small Groups [J]. Journal of Anthropological Research, 1976, 33(4): 452-473.
[26] GLEISER P M, DANON L. Community Structure in Jazz [J]. Advances in Complex Systems, 2003, 6(4): 565-573.
[27] GONG M, CAI Q, CHEN X, et al. Complex Network Clustering by Multiobjective Discrete Particle Swarm Optimization Based on Decomposition[J]. IEEE Transactions on Evolutionary Computation, 2014, 18(1): 82-97.

DONG Ming-gang, born in 1977, Ph.D. His main research interests include intelligent computing, multi-objective optimization and machine learning.

JING Chao, born in 1983, Ph. D. His main research interests include intelligent computing, optimization and deep reinforcement learning.

(上接第 456 页)

[6] MATATOV N, ROKACH L, MAIMON O. Privacy-preserving data mining: A feature set partitioning approach [J]. Information Sciences, 2010, 180(14): 2696-2720.
[7] GAERTNER W. A primer in social choice theory: Revised edition [M]. Oxford University Press, 2009.
[8] BRANDT F, CONITZER V, ENDRISS U, et al. Handbook of Computational Social Choice[R]. Cambridge University Press, 2016.
[9] ARROW K J. Social choice and individual values[M]. Yale University Press, 1963.
[10] KORNHAUSER L A, SAGER L G. Unpacking the court [J]. The Yale Law Journal, 1986, 96(1): 82-117.
[11] LIST C, POLAK B. Introduction to judgment aggregation[J]. Journal of economic theory, 2010, 145(2): 441-466.
[12] LANG J, PIGOZZI G, SLAVKOVIK M, et al. Judgment aggregation rules based on minimization[C]// Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge. ACM, 2011: 238-246.
[13] LANG J, SLAVKOVIK M. How hard is it to compute majority-preserving judgment aggregation rules[C]// Proceedings of the Twenty-first European Conference on Artificial Intelligence. IOS Press, 2014: 501-506.
[14] LANG J, PIGOZZI G, SLAVKOVIK M, et al. A partial taxonomy of judgment aggregation rules and their properties [J]. Social Choice and Welfare, 2017, 48(2): 327.
[15] LANG J, MENGIN J, XIA L. Voting on multi-issue domains with conditionally lexicographic preferences [J]. Artificial Intelligence, 2018, 265: 18-44.
[16] LI L, TANG X. A solution to the cold-start problem in recommender systems based on social choice theory [M]//Intelligent and Evolutionary Systems. Springer, Cham, 2016: 267-279.
[17] LI L I, NIU B, TANG S X. Online marketing research based on social choice theory and cloud computing[C]// 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC). IEEE, 2015: 391-396.

LI Li, born in 1982, Ph.D, lecturer. Her research interests include modern logic and artificial intelligence.