面向大数据的粒计算理论与方法研究进展

梁吉业[1,2], 钱宇华[1,2], 李德玉[1,2], 胡清华[3]
1. 山西大学计算智能与中文信息处理教育部重点实验室, 山西 太原 030006;
2. 山西大学计算机与信息技术学院, 山西 太原 030006; 3. 天津大学计算机科学与技术学院, 天津 300072

摘要
大数据的规模性、多模态性与增长性给传统的数据挖掘方法带来了挑战。粒计算作为智能信息处理领域中大规模复杂问题求解的有效方法，探索大数据分析的粒计算理论与方法有望为应对这些挑战提供新的思路和策略。瞄准若干大数据挖掘任务，对数据粒化、多粒度模式发现与融合、多粒度/跨粒度推理等方面取得的一些进展进行梳理和剖析，并针对天文数据挖掘和微博数据挖掘两个典型示范应用领域的初步研究进行了总结，以期为大数据挖掘领域的研究做出有益的探索。

关键词
大数据；粒计算；数据挖掘；信息粒化；多粒度

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016038

Research development on granular computing theory and method for big data

LIANG Jiye[1,2], QIAN Yuhua[1,2], Li Deyu[1,2], HU Qinghua[3]
1. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China
2. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
3. School of Computer Science and Technology, Tianjin University, Tianjin 300072, China

Abstract
Aiming at several data mining tasks, research developments on data granulation, multi-granularity pattern discovery and fusion, multi-granularity reasoning were carded and analyzed, and the preliminary study on two typical applications astronomical data mining and microblog data mining was summarized, which would be helpful for making a beneficial exploration in big data mining area.

Key words
big data, granular computing, data mining, information granulation, multi-granularity

1 引言

随着计算机技术、网络技术和传感器技术，特别是现代网络、云计算等技术的广泛应用，数据的生成和收集技术迅猛发展，数据量呈爆炸式增长态势，传统的数据处理技术遇到了极大挑战。在生物数据分析中，无论是DNA/RNA序列数据、蛋白质结构数据，还是代谢组数据、基因芯片数据，都是大数据中的典型类型数据。在社会媒体计算中，视频数据、语音数据、文本数据以及图像数据等都呈现出快速增长的趋势，如微博的用户量、访问时间以及微博信息量都快速增长。大数据在现代信息社会中的数据资源主体地位已成为学术界与企业界的共识，它不仅对经济活动与社会发展具有重要推动作用，也成为了世界主要经济体的战略研究计划。如何高效地从这些大数据中进行数据挖掘成为了当今信息科学领域研究的热点问题。

从大数据的外在来看，大数据经常呈现出大规模性、多模态性与增长性等特征，使得传统的数据分析理论、方法与技术面临可计算性、有效性与时效性等严峻挑战。

粒计算是专门研究基于粒结构的思维模式、问题求解方法、信息处理模式的理论、方法、技术和工具的学科，是当前智能信息处理领域中一种新的计算范式。通过分析大数据的表现形态、大数据挖掘面临的挑战与粒计算核心理念的内在关系可知，大数据自身具有天然的多层次/多粒度特性，数据挖掘任务也经常呈现多层次/多粒度特性，而大数据挖掘算法本身也要求可计算性、有效性、高效近似求解特性。这表明大数据的分析需求和粒计算框架有很强的契合性。

为了应对这些挑战，笔者着重在大数据的特征选择与信息粒化、多粒度模式发现与融合以及复杂决策任务的多粒度/跨粒度高效推理等方面做了初步研究，取得了一些重要的研究进展和成果。

2 大规模数据的特征选择与信息粒化

2.1 特征选择研究

（1）基于随机特征映射的降维方法

核方法是一类重要的机器学习方法，具有坚实的理论基础和完整的学习框架。该方法利用核函数将输入样本隐式地映射到一个高维甚至是无限维的特征空间，使原空间中的非线性可分样本在特征空间中实现线性可分。核矩阵保有核函数及输入样本的全部信息，是核方法表示和处理的核心。然而，核矩阵存储和计算的高复杂度成为核方法在大规模问题中的应用瓶颈。基于循环随机矩阵投影，笔者所在课题组提出了一种新的随机特征映射方法，将输入样本显式地映射到一个相对低维的随机特征空间，从而可以应用线性学习算法高效地求解非线性问题[1]。理论上，证明了提出的随机特征映射方法SCRF近似核函数是无偏的，并且较之前最先进的随机特征映射方法Fastfood具有更低方差。同时，提出的循环随机特征映射具有线性空间复杂度和对数线性时间复杂度低的特点，实现简单，是迄今最简单有效的随机特征方法。实验验证了SCRF的核估计准确率和计算效率，并且将SCRF应用于实际分类问题以验证该显式非线性降维方法的泛化性能。在核估计实验中，提出的SCRF计算的核估计值集中在对角线，表明提出的方法效果更好；并且在计算效率上，明显优于Fastfood。将精确核

方法（LIBSVM）和3种随机特征映射方法（RKS、Fastfood、SCRF）的测试准确率与效率（训练时间+测试时间）对比，可以发现3种随机特征映射方法相对于精确核方法得到了相当的测试准确率，很大程度提升了训练和测试效率。提出的SCRF的效率最为突出，并且随着训练数据的规模增加，效率提升越明显。因此，提出的循环随机特征映射给出了一种高效的非线性降维方法，并且具有坚实的理论基础，实现简单，能够广泛应用于大规模训练与预测问题。

（2）基于字典学习的特征选择方法

随着手持数字终端、工业传感器的广泛使用和社交网络的发展，大规模的、高维的、强不确定性的图像、视频、文本以及生物信息学数据大量涌现。高维的数据增加了存储负担，提高了算法的计算复杂度，降低了模型的泛化性能。因此降低特征空间维度、去除冗余和不相关的特征十分必要。

无监督特征选择的一个关键是如何生成伪的类标记刻画样本空间的数据分布，从而将无监督特征选择问题转化成一个有监督问题。目前主要采用谱分析、谱聚类、矩阵分解等方法生成伪的类标记，同时利用线性回归的模型学习得到一个特征选择矩阵。稀疏性也是特征选择的一个重要部分，通常特征选择矩阵会被要求具备组稀疏特性，以移除数据中的噪声特征。

笔者所在课题组首次将字典学习引入无监督特征选择任务中，提出了一种解析—合成字典对学习方法，通过表达系数刻画样本空间的数据分布[2]。合成字典用于重构样本，而解析字典通过一个线性投影把样本投射到表达系数空间。最后，解析字典用于评价特征重要性。本工作首次讨论了范数$L_{2,P}$对无监督特征选择的影响，并提出了利用迭代重加权最小二乘求解$L_{2,P}$范数优化问题，展示了不同P（范数）值下的无监督特征选择效果，证明了在$0<P<1$的情况下，提出的模型可以收敛到一个固定点。在标准的数据集上，验证了提出算法的性能优于目前的无监督特征选择算法。

（3）基于压缩表的符号数据特征选择方法

在符号型数据特征选择研究中，由于数据集每个特征下对象取值的数量较少，因此存在大量条件属性取值完全相同的对象，这使得数据中蕴含了大量的冗余信息。特别是，现有的特征选择算法对于这些相同的对象大多都是作为单独对象分别处理，这必然会导致大量的冗余计算，从而影响了特征选择算法的计算效率。为此，通过用一个对象代表与其特征取值相同的所有对象，提出了一种数据压缩表示方法，并在此基础上提出了基于压缩数据的特征选择算法[3]。

在理论分析方面，笔者所在课题组证明了在压缩后的数据集上得到的特征重要度与在原数据集上得到的特征重要度相等。由于特征重要度决定着每个特征被加入候选特征子集的顺序，进而决定了启发式特征选择的最终结果，因此从压缩数据和原始数据获得特征重要度一致的重要性质就确保了基于压缩后数据得到的特征选择结果与原数据集得到的特征选择结果完全一致。此外，课题组也在12个UCI公开数据集上对算法的有效性和效率进行了验证分析。从实验结果可以看出，与目前文献报道中最好的启发式特征选择算法（ACC-PR）相比，在大多数据集上基于压缩策略的特征选择算法（AR-CT-PR）可以在获得相同特征选择结果的同时，显著减少计算特征选择的时间消耗。这个策略在面向符号型大数据的处理时是一个重要、高效的分析策略，可在其他符号型数据分析任务中进一步推广使用。

2.2 信息粒化研究

（1）基于优化求解角度的符号数据聚类准则

聚类作为一类重要的信息粒化方法，不同的聚类算法或同一算法的不同参数设置往往在聚类同一数据时产生不同的结果。因此，人们需要聚类有效性函数去评测聚类结果，并从许多聚类结果中寻找最适合于数据的划分。面向数值型数据的聚类有效性评测方法已被人们广泛研究。但是，针对符号数据的聚类有效性评测研究相对较少。目前，针对符号数据，有3个广泛使用的有效性评测函数，其中包括：K-Modes目标函数F、分类效用函数CU和信息熵函数E。许多符号聚类算法以它们其中之一为聚类准则搜索聚类结果。当它们在数据聚类中被使用时，有以下3个问题需要解决。

- 它们在评测聚类结果上有怎样的共性和差异性。
- 当它们在评测聚类结果时类间信息是否被忽略。
- 以它们其中之一为聚类准则，如何确定该准则在一个数据集上的取值范围？

针对上述问题，课题组从解空间（优化）角度，系统研究了这3个有效性函数[4]。首先，构建了一个泛化的有效性函数及其优化模型。进一步，基于该泛化模型，分别对这些问题给出了理论性解释。

- 建立了这些有效性函数在评测聚类有效性上的内在关系，理论分析发现在评测聚类结果时，分类效用函数等效于信息熵函数，K-Modes目标函数的最优解是分类效用函数最优解的上界。
- 建立了这些有效性函数与类间评测函数之间的关系，理论分析发现最小化泛化函数等于最大化类间差异性。这暗示着使用这些类内信息评测聚类结果时并不会忽略类间信息。
- 对于一个给定的数据集，通过放宽某些变量的约束条件，将这些有效性函数最大化和最小化优化问题转化为凸规划问题，获得其上下界，从而帮助实现函数的归一化。

实验比较了来自UCI的12个数据集上的100次聚类结果的平均有效性。相比原始的有效性函数G，归一化后的函数的评测结果更接近于外部评测函数ARI和NMI。该研究成果为解决符号数据聚类准则的选择、聚类算法的互学习及数据特征对聚类有效性的影响等问题提供了理论基础。

（2）基于半监督的谱聚类的信息粒化

图像聚类在包含图像检索以及理解的实际应用中起着重要的作用。传统的图像聚类算法考虑单一的特征和固定的距离（如欧氏距离）来度量样本间的相似性。然而，不同的视觉特征往往能够提供互补信息对图像内容进行描述。此外，受限于时间和人力等的消耗，通常只获取到少量的标记样本，从而使得半监督学习成为一种必要的工具。为此，基于半监督距离学习和多模态信息，课题组提出了半监督的谱聚类算法对图像进行聚类[5]。通过提取颜色、纹理、形状以及语义等多种特征，利用少量的标记图像进行半监督距离学习，采用学习得到的度量以及高斯相似函数计算相似性，最终构造出半监督的拉普拉斯矩阵进行谱聚类。采用统计信息进行特征提取，可以对大小不同的图像进行聚类。大量实验结果表明，提出算法的性能优于传统方法。

（3）混合数据属性加权聚类的信息粒化

在传统的划分式聚类过程中，都假定各个属性对聚类的贡献程度相同，即在相似性或相异性度量的计算中所有属性的权重相同。而在大部分实际应用中，用户期望得到的聚类结果对参与聚类的各个属性的重要程度往往并不相同，特别是在高维数据聚类过程中，样本空间中各属性对聚类

效果贡献大小不同成为一个不可回避的问题。同时兼具数值型和分类型属性的混合数据在实际应用中普遍存在，混合数据的聚类分析越来越受到广泛的关注。

为解决高维混合数据聚类中属性加权问题，课题组提出了一种基于信息熵的混合数据属性加权聚类算法，以提升模式发现的效果[6]。工作主要包括：首先为了更加准确客观地度量对象与类之间的差异性，设计了针对混合数据的扩展欧氏距离；然后，在信息熵框架下利用类内信息熵和类间信息熵给出了聚类结果中类内抱团性及一个类与其余类分离度的统一度量机制，并基于此给出了一种属性重要性度量方法，进而设计了一种基于信息熵的属性加权混合数据聚类算法。在10个UCI数据集上的实验结果表明，提出的算法在4种聚类评价指标下优于传统的属性未加权聚类算法和已有的属性加权聚类算法，并通过统计显著性检验表明本文提出算法的聚类结果与已有算法聚类结果相比具有显著差异性。

2.3 多粒度空间的粒化不确定性

不同的信息粒化方法和策略将会导致给定数据的不同粒化结果，这意味着能够在这个粒度水平上观察和分析数据。认知主体在不同的粒度水平上观察的同一事物往往是不同的，它有一个所谓的粒结构来刻画。对于模糊信息粒化而言，模糊粒结构是一个数据集诱导的模糊信息粒的数学结构，模糊信息粒度则用于度量一个模糊粒结构的不确定性，也称为粒化不确定性。

为了有效地度量粒化不确定性，已经发展了若干形式的模糊信息粒度。然而，已有的模糊信息粒度度量有2个缺陷。一个是当两个模糊粒结构的信息粒度相等时，并不意味着它们是相等的，缺乏进一步区分模糊粒结构差异性的方法；另一个是目前的模糊信息粒度公理化方法仍然不够完备，不能够区分任意两个模糊粒结构的粗细程度。为此，课题组引进了一个所谓的模糊知识距离，用于刻画模糊粒结构之间的距离，理论分析表明它是一个距离测度，并且能够区分任意两个模糊粒结构之间的差异性；为了构造更加合理的模糊信息粒度公理化方法，基于提出的模糊知识距离提出了广义信息粒度公理化方法[7]，理论分析和实验结果都表明提出的这些新方法能够很好地刻画已有方法的以上两个不足，为模糊粒化不确定性研究提供了约束性框架。该研究为人类从不同角度、不同层次上认识大数据时采用的信息粒度水平提供了定量刻画方法，是面向大数据的粒计算理论与方法研究中的核心问题。

多粒度粗糙计算是通过多个粒化结构刻画目标概念，必然导致不确定性，该理论模型中存在知识粒和知识粒结构的不确定性，它直接决定问题求解的有效性。如何度量问题中的不确定性成为多粒度粗糙粒计算研究中的一个普遍问题。课题组借鉴了融合不确定性和不精确性的方法，提出融合信息熵、融合粗糙熵、融合信息粒度等度量，并讨论它们的重要性质，初步研究了多粒度近似空间中的不确定性[8]。这是针对多个粒空间诱导的粒化不确定性定量分析的首次尝试，将有助于多粒度空间的粒化不确定性的进一步研究。

3 大数据的多粒度模式发现与融合方法

3.1 基于联合概率估计的多模态信息融合

多模态数据分析核心问题之一是如何有效地进行多模态信息的融合。当前，针对

一些具体领域或任务已经开展了若干探索性研究，采用的主要策略是首先从不同模态数据中分别进行特征提取或特征选择，然后将提取出的特征合并成一个更大的特征空间，再按照传统的思路在此特征空间上进行数据挖掘。然而，这个策略可能会遇到不同变量之间语义不一致的问题。在数据挖掘等数据分析任务中，采用的分类、聚类、优化等方法都依赖于对象之间的某种距离测度，这需要将不同变量看作不同的维度并要求在这些变量上可进行线性运算。在视频分析中，通常可以从中提取出一些文本特征、图像特征、语音特征、场景特征等，尽管在特征向量化表示以后线性运算可以工作，然而在文本特征和图像特征之间进行线性运算的语义到底代表了什么，这些不同语义变量这样运算可能不是合理的，也许会影响最终的数据分析和挖掘效果。因此，如何克服不同模态特征之间的语义鸿沟是多模态数据挖掘的主要挑战之一。

为此，课题组提出一类较为一般的方法，将原始异构变量数据表转换为一种概率意义下的数据表，核心是将原来的距离测度转为任意两个对象是否相等的概率[9]。为了检验新的数据表示方法的有效性，首先在符号数据这种单一类型数据上进行了尝试。具体动机如下：目前最具代表性的符号聚类算法多数都是基于0~1距离或它的扩展版本来构造相似性测度，然而再反映到对象的簇结构中，由于这类距离不处在一个连续空间中，由它构造的相似性测度和基于频率的类中心更新可能不够有效。实验比较了最有代表性（聚类性能最好）的4种符号聚类算法以及笔者提出的SBC算法的两个版本在UCI的9个数据集上的100次聚类结果的平均聚类性能。从理论分析和实验结果可知，新的数据表示方案不仅保留了数据原始空间的簇结构，而且提供了更加丰富的测度信息。从中也可以看到，相比目前最具代表性的4种算法，提出的SBC算法在AC指标上平均有10%的提高，在ARI指标上平均有20%的提高。这表明提出的新数据表示方案有重要意义，为更加复杂的多模态数据分析提供了一种可资借鉴的有效方案。

3.2 基于深度神经网络的多模态特征融合与选择

深度学习是近年来兴起的一种有效的表示学习方法，已经在语音、图像等领域得到了成功的应用和长足的发展。借鉴深度学习的特征表示方法，课题组提出了一种结合深度神经网络与组稀疏方法的多模态特征选择算法，突破传统多源异构特征选择算法中存在的模态异构性带来的障碍，使用深度学习的方法对原始的异构多模态数据进行多重非线性变换，得到隐藏的抽象表达，将其从原始的异构特征空间转换到同一个特征空间之中[10]。进而使用Group LASSO的方法对这些同构特征进行选择，得到不同特征维度的权值，根据权值大小的不同选择出与当前给定学习任务最相关的特征维度用于最终的模式识别任务。具体地，给每一个模态都分配一个多层神经网络，从而形成一个多模态深度神经网络，用于将原始的异构特征转换为同一个语义层次的隐藏特征表达，得到同构的抽象特征。

同SVM（support vector machine，支持向量机）（使用所有原始异构特征）、MKL（multiple kernel learning，多核学习）（使用所有原始异构特征，为每一个模态分配一个核函数，使用多核学习方法进行融合核学习）、GLLR（logistic regression with group LASSO，基于组LASSO的逻辑斯特回归）（使用logistic regression with group LASSO 方法直接对原始异构特征进行选择）以及MMNN（multi-modal neural network，多模态

异构神经网络）（使用多模态异构神经网络得到的同构特征不做选择）等方法相比，提出的模型在3个实验数据集上训练SVM都取得了较好的分类效果，更是远远超过单独使用SVM分类器的分类精度。同时注意到本文算法在对模态进行选择后仍然取得了最高的分类精度，印证了多模态数据中信息冗余的存在与本模型滤除无关模态的有效性。

3.3 基于证据理论的多粒度融合方法

在现实世界中，多数据源指对相同数据样本采集于不同时间段或不同地方或是具有不同角度的数据描述。不同数据源的数据蕴含着数据样本中不同的结构信息，表达了数据样本间多种角度的信息。当同一数据样本的不同角度或者不同来源信息一起被使用时，数据样本之间蕴含的结构信息将更加丰富，这些结构信息在不同的应用中反映了学习任务的不同角度、不同侧面，要想全面理解数据中蕴含的多种信息，需要构造合理、有效的学习模型与算法。多源信息系统恰好可以用来表示这样的多源信息。因此多源信息的组合问题可以转化为多源信息系统的数据分析问题。从粒计算的角度来看，对每一个子信息系统，根据某种粒化策略生成对应的粒结构。换句话说，多源信息系统中来自不同源的信息可以看成不同的粒空间，从而多源信息融合问题也变为多粒空间融合问题。

课题组首次通过讨论经典多粒度粗糙计算模型与证据理论之间的联系，分别在清晰和模糊的两个多源背景下，讨论了乐观/悲观多粒度粗糙近似和证据理论的信任函数之间的关系，给出了多粒度粗糙近似空间证据的基本概率指派获取等问题[11]。借鉴K-Modes聚类的思想完成多个粒结构的聚类，结合证据理论，在多粒度视角下建立一类介于乐观融合和悲观融合之间的多粒度融合算法，称为基于证据理论的多粒度融合算法。并利用悲观模糊多粒度粗糙近似和模糊信任函数之间的关系，给出了粒度约简的理论框架。这些结果在一定程度上解决了多源不确定信息的定量和定性融合问题，也增强处理多源信息系统不确定问题求解的能力，为多粒度模式的知识发现奠定了一定的理论基础。

4 大规模复杂决策任务的多粒度/跨粒度高效推理模型和算法

4.1 多粒度单调分类器

单调分类（特征属性和决策属性存在单调性约束）是一类重要的分类任务。集成学习通过融合多个具有一定准确性和差异性的基学习器，能够大幅度地提高机器学习系统的泛化能力。然而，经典的集成学习方法通常都是通过改变原始训练数据集的分布得到不同的基分类器，然后对所有基分类器的输出进行简单投票得到最终的决策结果。基于改变样本分布的集成策略，通过在训练过程中提高分类器对不同数据的适应能力来降低预测方差，并没有从结构上产生具有差异性的分类器。

课题组基于粒计算的思想，利用特征属性和决策属性存在单调性关系的先验知识，在保序性约束的前提下，通过引入优势粗糙集，利用保持整体优势粒结构来寻找特征子空间，不同的子空间对应一个不同的粒结构；接着利用这些子空间来构造基分类器（个数可自适应确定）；然后利用最大概率原理对未知对象进行类别判别以实现多粒度分类器融合[12]。基于保序性得到的特征子空间能够在不同粒度下保持原始特征空间与决策属性之间的序结构信

息，从而保证了基分类器在单调分类任务中的个体性能。并且，在不同粒度下的保序性约束下，能够得到具有不同结构的特征子空间，从而得到具有结构差异性的基分类器。基于最大概率原理融合基分类器，综合了基分类器在每个类别上的性能优势，并且达到了基分类器之间互补的集成效果，相比投票方式利用了更多的决策信息。大量实验表明了多粒度分类器可极大提高单调分类任务的泛化能力，此外，集成使用的基分类器个数很少并且个数可自适应确定，大大降低了存储空间和预测时间。

4.2 基于层次结构的分类模型

物体的高层语义解释是图像识别中的关键问题。尽管机器学习算法在图像识别方面取得了很好的结果，但其效果远不如人的智能。这是因为人类识别物体发生在高层语义空间，而目前大多数机器学习方法仅仅通过底层的视觉特征对物体进行解释，这些方法虽然可以很好地描述图像的视觉内容，但不能像人类一样理解图像的高层语义。例如，一个人可能会把一条狼错误地分类成一条狗，却不会把一条狼错误地分类成一辆汽车。这是因为人类在分类时是以一种层次结构进行的，这种层次结构会把两个类之间的语义关系考虑进去，因此可以给出语义化的分类决策。利用层次结构分类会使得分类效果更准确，也更符合人类的语义认知。在不同视角下，类别之间表现出不同的类别语义关系，例如，在概念上基于词的语义关系和在特征上基于视觉相似性。

一些研究利用多种不同类别的语义相似关系，但是多种语义相似关系往往存在不同种相似关系的不一致性，例如在图像分类领域中，对比“鲸鱼”和“人类”两个类，“鲸鱼”与“人类”的视觉特征相似性远小于其概念相似性，因为在生物分类学中，鲸鱼和人类都属于哺乳动物，然而其视觉特征相差很远，这就会出现概念和视觉特征上的不一致问题。笔者提出了一种学习不同类别相似关系权值的方法，通过学习来融合得到最优的类别相似关系，由此构建类别间的层次结构[13]。对于层次分类问题，本工作将层次分类问题转化到在结构化学习框架下，利用结构化支持向量机进行分类，在两个图像数据集中验证了有效性。

5 示范应用研究

5.1 基于太阳观测大数据的空间天气预报建模

太阳动力学观测站（solar dynamics observatory，SDO）是美国宇航局（NASA）“与星同在”计划中发射的第一颗人造卫星，于2010年2月11日在美国亚特兰大成功发射，预计进行5~10年的观测任务，一直运转至今。它的目的是探究各种各样的太阳活动的成因及其对地球可能产生的影响。SDO总共搭载了3个科学实验仪器：大气成像仪（atmospheric imaging assembly，AIA）、日震与磁成像仪（helioseismic and magnetic imager，HMI）和极紫外线变化实验仪（extreme ultraviolet variability experiment，EVE）。其中，AIA利用多个紫外和极紫外波段，对太阳进行全日面、高时空分辨率的观测，平均每隔10 s在10个波段几乎同时对太阳进行成像；HMI分析太阳的磁场结构与活动以及太阳发生的变化；EVE拍摄太阳的极紫外线辐射，具有较高的光谱分辨率、时空分辨率和精确度。

SDO代表了太阳数据在数量和质量上新的前沿，它的成功发射，使得太阳物

理研究真正进入了大数据时代。每天拍摄150 000多张高分辨率的太阳图像(约1.5 TB),SDO任务生成的数据将超过之前所有太阳数据的总和。

(1)数据下载

使用洛克希德马丁太阳与天文物理实验室(Lockheed Martin Solar and Astrophysics Laboratory, LMSAL)研发的SSWIDL程序,通过国家天文台服务器,下载了2012年前6个月的太阳元数据(fits格式)。数据共包括9个AIA波段(分别为094、131、171、193、211、304、335、1 600、1 700)HMI磁动图,每个波段包括7 671张图像,每张图像为4 096×4 096分辨率。

(2)数据预处理

为了更好地可视化,将fit格式的原始灰度图像全部转化为JPEG格式的RGB图像。

(3)数据标注

根据太阳事件知识库(heliophysics event knowledgebase, HEK)提供的太阳事件报道信息,为每张图像生成标注文件(XML格式)。事件标注共包括6种太阳活动,分别为活动区(active region, AR)、冕洞(coronal hole, CH)、暗条(filament, FI)、耀斑(flare, FL)、西格玛型(sigmoid, SG)和黑子(sunspot, SS)及其对应的边界信息。

5.2 基于微博大数据的社会化推荐系统

5.2.1 资源建设

(1)汉语框架语义资源

课题组在山西大学汉语框架网(CFN)资源的基础上,新构建了67个框架,框架数量从304个增至371个,扩充了框架语义标注例句数量19 138条,词元4 585个,为支撑细粒度的文本语义分析需求提供了框架语义分析资源支撑。

(2)中文文本倾向性分析COAE2015微博语料库

构建了中文文本倾向性分析COAE2015微博语料库,涉及领域包括汽车、电子、手机、美食、娱乐、宾馆等,包括15 679条微博、20 154条观点句的标注及极性标注,并对13 787条观点句标注了24 093组细粒度观点要素及极性的三元组。另外,从新浪微博爬取的521个用户节点、4 936条关注关系以及每个用户发表的微博共计543 587条,为基于微博的社会化推荐系统提供了数据支持。

5.2.2 相关研究成果

课题组提出了一种基于细粒度篇章级框架语义分析的汉语阅读问答方法,给出了一种基于框架语义特征的文本零形式识别与填充方法[14],提出了基于相似性发现与训练数据调整的跨语言的文本情感倾向判别方法[15],建立了一类策略融合的跨语言文本情感倾向判别框架[16],发展了一种融合社交网络信息的协同过滤推荐算法[17]。

5.2.3 应用系统

研发了一个文本情感分析技术资源开放平台,主要包括微博数据的关键词抽取、观点要素抽取、文本情感分类以及基于汽车论坛和汽车口碑的汽车产品性能分析;研发了一个基于社交网络的好友推荐系统,包括用户模块、兴趣模块、展示模块以及其他附属模块四大功能模块,利用用户的微博内容和好友关系挖掘用户的兴趣偏好,个性化地为用户推荐相似程度高的潜在好友。

6 结束语

针对大数据的规模性、多模态性与增长性给传统的数据挖掘方法带来的挑战,

本文从粒计算的视角分析了应对这些挑战可能的新思路和新策略。具体面向数据的信息粒化、特征降维、多模态信息融合、特征学习与融合、多粒度证据融合、多粒度/跨粒度推理等问题，梳理和剖析了课题组取得的一些研究进展，并总结了在天文数据挖掘和微博数据挖掘两个典型示范应用领域方面的初步研究，以期为大数据挖掘领域的研究做出有益的探索。

参考文献:

[1] FENG C, HU Q H, LIAO S Z. Random feature mapping with signed circulant matrix projection[C]//The 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), July 25-31, 2015, Buenos Aires, Argentina. California: AAAI Press, 2015: 3490-3496.

[2] ZHU P F, HU Q H, ZHANG C Q, et al. Coupled dictionary learning for unsupervised feature selection[C]// AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. California: AAAI Press, 2016: 1-7.

[3] WEI W, WANG J H, LIANG J Y, et al. Compacted decision tables based attribute reduction[J]. Knowledge-Based Systems, 2015(86): 261-277.

[4] BAI L, LIANG J Y. Cluster validity functions for categorical data: a solution-space perspective[J]. Data Mining and Knowledge Discovery, 2015, 29(6): 1560-1597.

[5] LIANG J Q, HAN Y H, HU Q H. Semi-supervised image clustering with multi-modal information[J]. Multimedia Systems, 2016, 22(2): 149-160.

[6] 赵兴旺，梁吉业. 一种基于信息熵的混合数据属性加权聚类算法[J]. 计算机研究与发展，2016, 53(5): 1018-1028.
ZHAO X W, LIANG J Y. An attribute weighted clustering algorithm for mixed data based on information entropy[J]. Computer Research and Development, 2016, 53(5): 1018-1028.

[7] QIAN Y H, LIANG J Y, DANG C Y. Fuzzy granular structure distance[J]. IEEE Transactions on Fuzzy Systems, 2015, 23(6): 2245-2259.

[8] LIN G P, LIANG J Y, QIAN Y H. Uncertainty measures for multigranulation approximation space[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2015, 23(3): 443-457.

[9] QIAN Y H, LI F J, LIANG J Y, et al. Space structure and clustering of categorical data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015: 1-13.

[10] ZHAO L, HU Q H, WANG W W. Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso[J]. IEEE Transactions on Multimedia, 2015, 17 (11): 1936-1948.

[11] LIN G P, LIANG J Y, QIAN Y H. An information fusion approach by combining multigranulation rough sets and evidence theory[J]. Information Sciences, 2015, 314(1): 184-199.

[12] QIAN Y H, XU H, LIANG J Y, et al. Fusing monotonic decision trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(10): 2717-2728.

[13] ZHAO S, ZOU Q. Fusing multiple hierarchies for semantic hierarchical classification[C]//The 8th International Conference on Machine Learning and Computing, February 22-23, Hong Kong, China. [S.l.:s.n.], 2016: 47-51.

[14] LI R, WU J, WANG Z Q, et al. Implicit role linking on Chinese discourse: exploiting explicit roles and frame-to-frame relations[C]// The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural

Language Processing, July 26-31, 2015, Beijing, China. [S.l.:s.n.], 2015: 1263-1271.

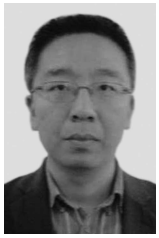
[15] ZHANG P, WANG S G, LI D Y. Cross-lingual sentiment classification: similarity discovery plus training data adjustment[J]. Knowledge-Based Systems, 2016, 107(1): 129-141.

[16] 张鹏, 王素格, 李德玉.一种策略融合的跨语言文本情感倾向判别方法[J]. 中文信息学报, 2016, 30(2): 32-40.
ZHANG P, WANG S G, LI D Y. A multi-strategy approach to cross-lingual sentiment analysis[J]. Journal of Chinese Information Processing, 2016, 30(2): 32-40.

[17] 郭兰杰, 梁吉业, 赵兴旺. 融合社交网络信息的协同过滤推荐算法[J]. 模式识别与人工智能, 2016, 29(3): 281-288.
GUO L J, LIANG J Y, ZHAO X W. Collaborative filtering recommendation algorithm incorporating social network information[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(3): 281-288.

作者简介

梁吉业(1962-),男,博士,山西大学计算智能与中文信息处理教育部重点实验室、山西大学计算机与信息技术学院教授,主要研究方向为人工智能、粒计算、数据挖掘与机器学习。



钱宇华(1976-),男,博士,山西大学计算智能与中文信息处理教育部重点实验室、山西大学计算机与信息技术学院教授,主要研究方向为人工智能、数据挖掘与机器学习。

李德玉(1965-),男,博士,山西大学计算智能与中文信息处理教育部重点实验室、山西大学计算机与信息技术学院教授,主要研究方向为数据挖掘与机器学习、粒计算、概念格。

胡清华(1976-),男,博士,天津大学计算机科学与技术学院教授,主要研究方向为人工智能、机器学习、模式识别。

收稿日期: 2016-06-20
基金项目: 国家自然科学基金资助项目(No.61432011, No.U1435212)
Foundation Items: The National Natural Science Foundation of China(No. 61432011, No. U1435212)