

大数据挖掘面临的挑战与思考

梁吉业

本报告主要包括大数据处理背景、大数据挖掘面临的几个问题和基于粒计算的大数据挖掘算法3部分内容。

1 大数据处理背景

大数据大致可分为科学大数据、Web大数据和决策大数据,其中科学大数据是指科学实验、科学观测、科学设计等科学研究领域产生的数据;Web大数据是由门户网站、社交网络、搜索引擎、电子商务等领域产生的数据;决策大数据是由传统数据库和数据仓库在生产过程中产生并用于决策的数据。大数据挖掘呈现出一个倒立的金字塔型,底层是IT计算系统和平台层,中间是算法和模型层,顶层是应用层。目前大数据挖掘更强调数据的整体性、相关性、问题的满意近似解和深度分析与挖掘。近年来,大数据的发展也进一步推动了人工智能的发展,比如谷歌AlphaGo的围棋之战、微软小冰的人机对话和谷歌的无人驾驶汽车。随着人们对大数据挖掘认识的不断深入,不要一味追求“数据规模大”,而应更关注数据的整合与融合;不要一味追求结果的精确性,而应注重时效性、精确性以及完备性三者之间的关系,这成为大数据挖掘的新共识。

2 大数据挖掘面临的几个问题

(1)大数据的超大规模性和快速增长性给传统的聚类分析方法带来了巨大的计算困难,如何设计快速有效的聚类算法?

大数据的分解与融合是解决大数据超大规模性和快速增长性的一种有效方法,通过大数据的分解可产生相应的基聚类器,通过融合策略实现聚类的集成。我们重点关注两个问题:①基聚类器集成问题,主要任务是以某个强聚类器为参照物,将多个弱聚类器进行融合,形成强聚类器。核心问题是如何提取集成方向,构建集成模型,实现“弱弱生强”。②局部聚类结果集成问题,主要任务是将多个局部结果进行融合,逼近全局聚类结果。核心问题是聚类器不在整个数据集上运行的情况下,如何获得全局聚类结果的先验信息。

(2)大数据时代很难获得大量的有效标签,数据呈现明显的无监督或极弱监督特性,如何发展有效的分类方法?

传统分类学习方法通常需要事先标注(一部分)对象的类别标签,通过在训练集上学习分类器,并在测试集上检验其性能。对于大数据分类问题而言,标注一定比例的标签是费时费力的,甚至是不可能的。大数据分类学习任务可能获得标记的对象仅仅占整个大数据集非常微小的一部分,呈现明显的极弱监督性。重点关注的问题有:①极弱监督分类学习,主要任务是利用无标记样本去改善在小数据集上构建的分类器,核心问题是如何保证监督学习与无监督学习对“类假设”的一致性。②极弱监督聚类学习,主要任务是利用极弱的监督信息,改善无监督聚类结果。核心问题是如何利用无标记样本,增强极弱监督信息。③主动+半监督分类学习,主要任务是迭代选取无标签样本中的少量核心数据,并进行人工标注,构造半监督分类器。核心问题是如何度量无监督样本的重要性以及如何确定人工标注数据的停止准则。

(3)大数据中不同数据集中的特征呈现出复杂的关联纠缠状态,如何有效地挖掘数据中隐含的关联关系?

关联性分析是指发现存在于大量数据集中的关联关系或相关关系,从而描述事物中某些属性同现的规律和模式。这种同现关系可能表现为具有严格确定性的函数显示表达形式,也可能是客观对象之间确实存在但在数量上不是严格对应的依存关系,也可能是不存在内在联系的伪相关关系。现有的相关性分析可分为统计相关分析、互信息、矩阵计算和距离。关联分析的代表性工作有偏向于线性相关关系的Person相关系数和可以识别任何形式关联关系的最大信息系数(MIC)。在大数据环境下,针对关联挖掘可能采取的策略包括:①构造性方法,对于有特定相关性约束的关联关系,可构造出具体度量。②学习性方法:利用人类联想、关联的认知机理,从已有的关联关系中学习和识别新的关联关系。③探索性方法,对于未知的复杂关联关系,可通过探索性策略来挖掘。

3 基于粒计算的大数据挖掘算法

粒计算是研究基于信息粒及其结构的问题求解、信息处理的理论、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。大数据挖掘的特点与粒计算范式具有高度契合性。IEEE SMC 前主席 C. L. Philip Chen 教授于 2014 年在《Information Sciences》杂志上发表了大数据综述性论文,对大数据处理的潜在技术和研究进行了展望,其中将粒计算作为大数据分析挖掘的一种重要方法。针对大数据的规模性、多样性和增长性,我们从动态粒度和多粒度视角分别提出了基于样本表征整体的特征选择算法、基于批增量的特征选择算法和复杂数据的有效聚类算法。

进一步研究的几个问题:

- (1)针对数据的多样性,开展数据粒化、特征粒化和多粒度融合的研究,以提高问题求解的效率。
- (2)探索融合无监督深度学习和无监督概念学习的新型粒计算模型,克服无监督深度学习对数据规模的饥渴挑战和无监督概念学习对先验知识的饥渴挑战。
- (3)探索基于多粒结构的多模态变量的关联关系挖掘,为复杂数据关联关系分析提供有效途径。

本文根据 2016 年第六届中国数据挖掘会议(CCDM2016)大会报告“大数据挖掘面临的挑战与思考”整理。

梁吉业 教授,博士生导师,山西大学计算智能与中文信息处理教育部重点实验室主任,太原师范学院院长,主要研究方向为人工智能、数据挖掘与机器学习、智能管理与决策等。