北京邮电大学

硕士研究生学位论文开题报告

学 号: 2015140884

姓 名: 应臻奕

学 院: 网络技术研究院

专业(领域): 计算机技术

研究方向: 大数据计算和分析

导师姓名: 高志鹏

攻 读 学 位: 工程硕士

2016年 11月 25日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 基于AP聚类的不完备数据处理的研究与实现 | | |
| 选题来源 | 中央、国家各部门项目 | 论文类型 | 应用研究 |
| 开题日期 | 2016-12-8 | 开题地点 | 新科研楼509 |
| **一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）**  信息技术的快速发展，孕育了大数据时代的到来。大数据已然成为信息社会的重要财富，为人们更深入地感知、认识和利用物理世界提供了前所未有的丰富信息。而把信息变为知识的核心，在于数据挖掘。数据挖掘可以分为目标设计、预处理、数据挖掘、模式评估、知识表示等5个步骤，其中预处理占了整个工作的60%。[1]不完备数据处理作为预处理的一部分，一直都是研究的热点。  随着信息的膨胀，数据规模的不断扩大，各种数据质量问题混杂其中，降低了数据的可用性，其中数据缺失问题更是难以避免。不正确的度量方法、收集条件的限制、手动录入时出现遗漏或者主观有意隐瞒等问题都可能导致数据中出现大量“空值”。[2]如果忽略这些 “空值”，则会影响后续的数据挖掘、统计分析等工作的进行；如果将存在“空值”的数据元组删除，则会丢失大量的原始数据信息；由此可见，填补是相对科学的处理方法，但是如果对缺失值只是进行简单的常值填补，则忽略了缺失值的偏差，会改变样本的分布[3]。综上，由于数据的不完备问题不可避免且极大地影响了数据挖掘的难度和质量，因此对不完备数据如何进行有效的填补是我们亟待解决的问题。为此，本课题计划研究不完备数据的处理方法。  目前国内外提出了多种对不完备数据的填补方法，大致可分为两类：统计学方法[4]和机器学习方法[5]。统计学方法主要有回归法[6]、EM算法[7]、多重插补[8]等，大多利用整个数据集对缺失数据进行填补，没有充分考虑到数据对象的类别特征，使得填补值容易受到不同类别对象的干扰，降低填补结果的准确性。[9]机器学习方法，是随着研究热潮慢慢应用在缺失值处理中，是对不完备数据进行分类或聚类从而进行填补，主要有决策树[10]、KNN[11]、贝叶斯网络[12]、神经网络[13]、K-means[14]等方法。分类方法将含缺失属性作为目标属性，普遍存在一个缺失属性一个模型的现象，当含缺失属性较多时效率较低；而聚类方法将相似的对象聚成一个类，依据簇中相似对象进行填补，并不受缺失属性的影响，有更普遍的适用性。  AP聚类，是Frey和Dueck在2007年发表于《Science》杂志的聚类算法。算法将所有数据点作为潜在的聚类中心，在相似度矩阵的基础上通过消息传递进行聚类。[15]与传统聚类相比，AP聚类的相似性矩阵可采用任意度量方法，不需要满足三角不等式或距离对称等约束条件，也不需要事先指定聚类中心和聚类个数，同时算法聚类结果稳定。[16]本课题拟利用AP聚类的这些优点，将其有效地应用于不完备数据的处理过程中，提高缺失值填补的准确度。  主要参考文献：  [1] Cios K J, Kurgan L A. Trends in Data Mining and Knowledge Discovery[M]// Advanced Techniques in Knowledge Discovery and Data Mining. Springer London, 2005:1-26.  [2] 金连. 不完全数据中缺失值填充关键技术研究[D]. 哈尔滨工业大学, 2013.  [3] Cismondi F, Fialho A S, Vieira S M, et al. Missing data in medical databases: Impute, delete or classify?[J]. Artificial Intelligence in Medicine, 2013, 58(1):63-72.  [4] 曹林. 基于统计学习的数据预处理缺失值清洗方法研究[D]. 哈尔滨工程大学, 2012.  [5] Jerez J M, Molina I, García-Laencina P J, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem[J]. Artificial Intelligence in Medicine, 2010, 50(2):105-115.  [6] 汪静波. Logistic回归模型中缺失数据的处理[D]. 南京大学, 2015.  [7] Rahman M G, Islam M Z. Missing value imputation using a fuzzy clustering-based EM approach[J]. Knowledge and Information Systems, 2016, 46(2):389-422.  [8] Lee K J, Simpson J A. Introduction to multiple imputation for dealing with missing data.[J]. Respirology, 2014, 19(2):162–167.  [9] 冷泳林, 陈志奎, 张清辰,等. 不完整大数据的分布式聚类填充算法[J]. 计算机工程, 2015, 41(5):19-25.  [10] 高丽君. 面向缺失数据的变精度粗糙集决策树分类算法研究[D]. 大连海事大学, 2013.  [11] Huang J, Sun H. Grey Relational Analysis based k Nearest Neighbor Missing Data Imputation for Software Quality Datasets[C]// IEEE International Conference on Software Quality, Reliability & Security. 2016.  [12] Niloofar P, Ganjali M. A new multivariate imputation method based on Bayesian networks[J]. Journal of Applied Statistics, 2014, 41(3):501-518.  [13] Gautam C, Ravi V. Data imputation via evolutionary computation, clustering and a neural network[J]. Neurocomputing, 2015, 156(C):134-142.  [14] Patil B M, Joshi R C, Toshniwal D. Missing Value Imputation Based on K-Mean Clustering with Weighted Distance[M]// Contemporary Computing. Springer Berlin Heidelberg, 2010:600-609.  [15] Frey B J, Dueck D. Clustering by passing messages between data points.[J]. Science, 2007, 315(5814):972-6.  [16] 刘晓勇, 付辉. 一种快速AP聚类算法[J]. 山东大学学报(工学版), 2011, 41(4):20-23. | | | |

|  |
| --- |
| **二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**  （一）研究内容  本课题根据不完备数据处理的步骤，主要包含了聚类和填补两个方面的研究。  （1）对不完备数据的增量式AP聚类  对不完备数据的增量式AP聚类的研究内容，包含三个部分：对不完备数据的AP聚类、增量式AP聚类和算法的分布式化。   1. 对不完备数据的AP聚类   传统的聚类方法，像K-means、DBSCAN，依据距离公式来比较两个点的相似度，其度量方法比较固定，大多适用于完备数据集的聚类，在带有缺失数据的不完备数据集中，无法直接测量数据对象之间的距离，而无法直接应用聚类算法对不完备数据集进行聚类。而针对这个问题，本课题拟采用AP聚类算法对不完备数据进行聚类。AP聚类算法在相似度矩阵的基础上通过消息传递进行聚类。由于AP聚类的相似性度量可以采用任意度量方法，AP聚类可以有效地适应多样的数据，包括不完备数据。  本课题拟研究一种新的针对不完备信息系统的相似性度量方法，通过这种方法可以度量不完备数据对象间和不完备数据与完备数据对象间的相似度，进而利用 AP聚类算法对不完备数据集进行聚类。   1. 增量式AP聚类   根据不完备数据是否参与聚类，传统的聚类填补方法可以分为两种。一是直接对不完备数据集进行聚类，或对不完备数据进行简单填补后进行聚类，根据不完备数据所属的簇进行填补。这种方法，考虑了不完备数据在数据集中的影响，对不完备数据的聚类处理比较粗糙，增加了聚类的难度，同时由于不完备数据的信息干扰，聚类结果准确度较低。二是对完备数据集进行聚类，再将不完备数据与各个聚类中心进行相似性比较，根据相似性最高的簇中心作为填补依据。这种方法完全依靠完备数据集进行聚类，直接除去了不完备数据参与聚类的过程，忽略了不完备数据作为数据集一部分的影响。  本课题拟分别参考这两个方法的优缺点，提出对不完备数据进行增量式聚类的方法。考虑不完备数据加入聚类过程，但是由于不完备数据的信息不完整，先由完备数据集进行聚类，同时将不完备数据依照所含信息量进行排序，以增量方式对不完备数据集，依信息量由高到低依次聚类，在增量式聚类的同时进行数据填补，这样来保证数据集信息的最大化利用，提高填补的准确度。   1. 算法的分布式化   AP 聚类分析本身是迭代学习的过程，需要多次迭代直至聚类结果收敛。当应对互联网、科学计算领域海量、高维度的数据时，大数据量的计算及迭代给聚类分析应用带来了沉重的计算负载，使得计算时间过长，响应时间缓慢。传统的聚类算法因单机CPU、内存等资源的限制执行效率根本不能满足大数据处理的计算需求。因此，研究算法的并行化，并行的处理数据，可以有效的提升算法的执行效率，缩短数据处理的时间。  以Hadoop、Spark为代表的分布式云计算平台，充分利用多台计算机的存储资源与计算能力，提供简易的编程接口，方便实现聚类算法的分布式并行。基于分布式平台的可扩展性和容错性，聚类算法能极大满足数据分析时的伸缩性和可扩展性。本课题拟将对不完备数据的增量式AP聚类分布式化，使算法可扩展，依托分布式系统提高处理效率。  （2）改进KNNI的填补  在聚类之后，最简单的缺失属性填补方法，是采用聚类中心的值进行填补。但是尽管簇中心能从总体上一定程度地代表簇，但簇中心的个别属性并不能代表簇中其他对象的属性。另外由于存在边界问题，含缺失数据对象可能处在簇的边界，这时候取簇中心的值来填补缺失属性，会降低准确度。还有，基于平均值的填补方法。这种方法只参考了簇中缺失属性的平均值，没有考察实例的其它属性，同样会丢失一些有用的信息，降低填补准确度。另外，利用关联规则对缺失值进行填补也是一个很好的方法，但在实践中，通常会遇到两个问题：第一，产生的规则太少，不足以对所有的缺失值进行填补；第二，存在多条规则可以对同一缺失值进行填补，但填补值不相同。  KNNI通过用k个近邻值的平均值来代替缺失值，这个均值是带有权重的，权重与被预测点的距离呈反向变化。KNNI本身可以作为一个优秀的填补算法，但填充效果很大程度取决于数据的质量。在现实世界中，数据中的噪声是不可避免的，当 k个最近邻存在噪声时，用来填充就会产生较大的偏差。而在聚类之后，来自其他簇的噪声就被过滤了，可以很好地进行缺失属性填补。还有一个问题是，k值的确定比较困难。针对k值的确定，当下也有不少改进的研究。例如，QENNI，使用缺失数据象限方向的最近邻数据填充该缺失值 ,避免了KNNI中选取的k个最近邻点有偏好的情况；KNNI-SC利用稀疏编码理论和重构技术来解决KNNI算法关于参数 K 的选取问题。这些方法都是在特定的数据问题上有很好的改进效果。  本课题拟改进KNNI的算法，提出一种递进式的最近邻填补。通过改变权重的公式，使相似度高的数据对象有一个较高的比重，相似度低的数据对象的比重能很快的趋向于0，将其表现在填补值上，使填补值快速收敛，当填补值基本不变时，停止加入最近邻对象。这样不需要确定最近邻的k值，同时能很好地利用含缺失对象的最近邻信息。  （二）研究目标和效果  本课题的研究根据研究内容需要达到以下的目标。  （1）针对不完备数据的AP增量式聚类  1. 设计提出一种新的针对不完备信息系统的相似性度量方法，通过这种方法可以度量不完备数据对象间、完备数据对象间和不完备数据与完备数据对象间的相似度，能很好地与AP聚类相结合，使AP聚类适用于不完备信息系统的处理。  2. 设计提出一种针对不完备数据的增量式聚类的方法，通过这种方法可以充分有效地利用不完备数据集中完备部分与不完备部分的数据信息，实现不完备信息系统的递进式学习，提高聚类算法的准确度。  3. 实现AP增量式聚类算法的分布式化，通过将算法移植到分布式系统，利用分布式系统的并行处理，来提高算法的效率，同时使算法能在分布式系统上有很好的可扩展性。  （2）改进KNNI的填补  设计提出一种递进式的最近邻填补方法。通过这种方法，不需要额外的k值的设定，利用相似度来定义它的权重，从而实现缺失填补值的有效收敛，同时保证填补的准确度。  （三）拟解决的关键科学问题  课题的研究内容在解决的过程有以下几个关键问题。  （1）针对不完备数据的增量式AP聚类  1．在相似度度量过程中，如何有针对性的对待完备数据对象间、不完备数据对象间、完备数据对象和不完备数据对象间的处理，包括不同属性的处理，例如数值型属性以及分类型属性等，使度量方法能很好地应用于AP聚类。  2．在将AP聚类算法增量式化过程中，考虑AP聚类算法的复杂度，如何合理地安排增量式过程的分配，在考虑算法效率问题的同时，使含缺失数据对象能合理地加入聚类过程，尽可能多地利用数据集完备部分和不完备部分的有效信息，实现算法增量式的学习。  3. 在将算法分布式化过程中，如何合理有效地将数据分块处理再合并以及算法的并行化计算，使算法能有较好的扩展性，尽可能提升算法的效率。  （2）改进KNNI的填补  1. 在权重公式的设计上，在不降低填补值的准确度前提下，如何保证权重随着相似度的降低快速趋向于0，避免震荡，从而使填补值能快速收敛。 |

|  |
| --- |
| **三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**  本课题的研究整体拟采用文献调研、问题分析、提出解决方案、实验验证的研究方法，下面详述各部分拟采用的技术路线，以及可行性分析。  （一）技术路线  （1）针对不完备数据的增量式AP聚类  1. 本课题会对目前已有的针对不完备数据的相似度度量方法进行文献的查找与整理，学习并比较各个算法对不完备数据的相似度度量的优缺点，结合不同度量的优点，设计一种新的针对不完备数据的相似度度量方法，并应用于AP聚类。AP聚类的相似性度量没有方法的限制。通过实验验证来证明其在相似度度量上的有效性，及对于AP聚类的适用性。  2. 本课题会对目前已有的增量式聚类算法进行文献的查找与整理，学习各个聚类算法如何与增量式学习过程有效的结合，设计一种针对AP聚类的增量式学习算法。结合不完备数据的相似度度量方法，对不完备数据集进行增量式AP聚类。通过实验处理不同数据集，测量算法的准确度，来验证算法的有效性和鲁棒性。  3. 本课题会对目前已有的算法分布式进行文献的查找与整理，学习算法分布式化过程中的数据分块处理和并行化计算，代码编写实现针对不完备数据的增量式AP聚类算法在分布式系统上的移植。通过实验验证，测量算法的运行时间、加速比和吞吐量，证明算法的可扩展性和效率的提升。  （2）改进KNNI的填补  本课题会对KNN以及KNNI等相关改进算法进行文献的查找与整理，学习并比较各个改进算法的优缺点，并设计一种新的KNNI的改进。结合AP聚类，针对不同缺失数据集进行处理，测量算法的准确度，用实验验证改进算法的有效性和鲁棒性。  （二）可行性分析  （1）针对不完备数据的增量式AP聚类  AP聚类算法是Frey和Dueck在2007年发表于《Science》杂志的聚类算法。自发表之后，AP聚类算法的优良表现使其获得了广泛的关注和研究。到目前为止国内外学者在AP聚类的研究上已经有了较多的理论和实践成果可供参考，为本课题的研究奠定了坚实的理论背景。本课题拟通过改进AP聚类算法使其应用于不完备数据的处理。当前也已有不少的相关研究和成果可以借鉴，因此本课题研究方法可行。  （2）改进KNNI的填补  KNN算法是比较经典的分类算法。国内外学者在KNN算法的研究与应用上已经有了相当多的理论和实践成果，为本课题的研究奠定了坚实的理论基础。本课题拟对KNN算法进行关于K值确定问题的改进，当前也有不少相关的研究和成果可以借鉴，包括基于象限的KNN填补和基于稀疏编码理论的KNN填补等，因此本课题研究方法可行。 |

|  |
| --- |
| **四、本研究课题可能的创新之处（不少于500字）**  本课题的创新点包括如下几点：  （1）目前利用聚类填补的方法主要是两种。一是，只针对不完备数据中完备部分数据来进行聚类，将不完备数据对象与聚类中心进行相似度比较，从而将不完备数据对象划分到各个聚类的簇中，再根据簇中信息来填补。这种方法在聚类时忽略了不完备数据的信息，没有考虑数据集的整体分布情况，降低了聚类的准确度。二是，对不完备数据进行简单填补或不处理，直接进行聚类，根据不完备数据对象所属的簇进行填补。这种方法未重视不完备数据带来的信息干扰，影响聚类的准确度，同时增加了聚类的难度。这两种方法，对于不完备数据在聚类过程中的处理都比较粗糙，没有充分利用数据集的信息，容易导致聚类的不准确。本课题拟提出一种增量式的聚类算法，将数据集中的完备数据进行聚类，形成聚类的初步框架，再将不完备数据进行增量式的聚类，充分利用数据集的信息，提高聚类的准确度  （2）目前KNN填补算法主要有两个问题。一是，容易受不同类别的对象干扰。将KNN填补算法与AP聚类相结合能很好避免此类问题。二是，K值的确定较困难，因为不同缺失对象周围近邻对象的分布不同。固定K值，填补的效果会随着对缺失对象分布的不同而改变；变化K值，很难依照分布自适应地变化。本课题拟提出一种递进式的KNN填补算法，缺失对象受最近邻的影响最高，相似度越低，影响越小，将这个影响表现在填补值上，使填补值快速收敛。这样不需要额外设定K值，更适应普遍情况。  （3）目前的聚类填补算法基本适应于单机器处理。当遇到海量数据时，在单机上进行聚类填补会遇到内存容量和内核处理速度的瓶颈问题。本课题拟将算法移植到分布式系统上，利用分布式计算框架对算法进行并行化改进，实现算法的可扩展，从而大幅度提高聚类填补算法的效率。 |
| **五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）（不少于500字）**  （一）与本项目相关的研究工作积累基础  （1）在数据挖掘方面，作者之前参与了“面向电商应用的大数据分析平台研发及应用”的大数据分析项目。该项目主要是基于Hadoop平台对电商数据进行数据挖掘并可视化呈现。在此过程中，作者学习并实践了机器学习的一些经典算法，如朴素贝叶斯、决策树等，对数据挖掘有了比较深刻的认识。  （2）在分布式系统方面，作者自行学习了Hadoop并在虚拟机上进行了搭建，同时使用了MapReduce框架进行了一些简单的分布式处理，初步认识了并行化计算框架。这些工作对于本课题进行分布式的移植有了很好的实践基础。  （3）此外，在不完备数据处理的研究上，已经进行了大量的前期调研工作，阅读了大量统计学方法和机器学习方法相关的文献，并学习了多重插补、关联规则、贝叶斯网络等相关知识，为后期的研究奠定了坚实的理论基础。  （4）掌握论文撰写方法，具备一定的论文阅读总结撰写能力和实验验证的技能。  （二）已具备的实验条件以及尚缺少的实验条件和拟解决途径  已具备的实验条件：  （1）已获取来自UCI的公开数据集。  （2）自行编写了KNNI算法。  （3）实验室有4台Centos 6.5的刀片服务器，已部署Hadoop/Spark集群。  还缺少的实验条件：  （1）AP聚类算法的实现代码。  （2）含不同缺失的人工数据。  拟解决的途径：  （1）查阅相关文献，仔细研读算法改进部分。  （2）网上查找或自行编写AP聚类算法的代码。 |

**学位论文工作计划**

|  |  |  |
| --- | --- | --- |
| 时间 | 研究内容 | 预期效果 |
| 2016.9-2016.12 | 不完备数据处理方法的前期调研 | 充分地研究现有的不完备数据处理方法，为研究奠定理论基础。 |
| 2017.1-2017.5 | 不完备数据处理方法的详细设计 | 设计出基于AP聚类的不完备数据处理方法的具体方案，满足课题预设的目标。 |
| 2017.6-2017.8 | 代码实现设计的不完备数据处理方法 | 能够结合已有的分布式平台，对设计的不完备数据处理方法进行验证。 |
| 2017.9-2017.10 | 实验验证及方案的优化 | 验证设计的方案，并对比其他研究者的方案，如可能，做进行进一步优化。 |
| 2017.11-2018.2 | 撰写论文 | 完成毕业论文的文字总结。 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 评  定  小  组  成  员 | 姓 名 | 职 称 | 单位名称 | 职务 |
|  |  |  | 组长 |
|  |  |  | 成员 |
|  |  |  | 成员 |
|  |  |  | 成员 |
|  |  |  | 成员 |
|  |  |  | 成员 |
| 导师意见： | | | | |
|  | | | | |
| 导师（签名）：  日期： 年 月 日 | | | | |
| 开题报告小组意见： | | | | |
| 组长（签名）：  日期： 年 月 日 | | | | |
| 学院意见（签章）： | | | | |
| 负责人：  日期： 年 月 日 | | | | |