十年前六个数据挖掘算法

张师超

广西师范大学数据挖掘实验室

2017.5.7　　桂林

作者简介

张师超教授是国家千人计划创新人才第五批获得者，人事部首批海外高层次留学人才回国资助人选获得者，中国科学院计算所兼职研究员。长期从事数据挖掘和大数据的研究，取得了一系列有重要价值的研究成果。在 Springer 出版学术专著 3 部，重要国际期刊独著论文 12 篇、一作或通信作者国际核心国学术期刊发表学术论文 70 余篇、国际顶级会议论文 10 多篇。据 Google Scholar 检索，这些成果获得引用 5000 多次，作者在 2014、2015 和 2016 年连续进入了 Elsevier 发布的中国高被引学者榜单。先后主持了国家级项目共 13 项，获得国际会议最佳论文奖 2 项，入选广西“八桂学者”、广西高校首批跨世纪人才培养对象。曾经应邀到新加坡、澳大利亚等的大学讲学或合作研究 10 多所/次。历任《知识与数据工程 IEEE 会报》、《知识与信息系统》和《智能信息学 IEEE 通报》的副编辑，国际会议主席或者程序主席/副主席 9 次。

前言

在指导硕士和博士研究生的 20 余年中，一直讲授我自己对为人与做学问的理解方法，效果还不错。按照原计划，我打算通过出版社出版自己认为比较满意的这些论文，供潜心研究的同行和应用开发者鉴享。联系后才发现，需要的时间很长。于是，我就有了在网站进行交流和讨论的想法。

在我 30 多年的研究生涯中，前十年主要研究时态推理与不确定性推理，后二十多年研究数据挖掘。考虑到版权的问题，这个精选论文集基本上是选择发表了十年或以上的论文。的确，这肯定有过时的嫌疑。但是，我相信，论文中的思路和解决方法依然有一定的参考价值。这是因为，近几年来发表大数据方面论文的国际同行对选于本论文集中的一些方法予以肯定，鼓舞了我向读者推荐这组论文的信心。

这组论文包括数据划分挖掘、数据挖掘、负关联规则挖掘、时态区间演算和不确定性推理的矩阵计算，共 5 个方面的研究成果。为了忠实原文，直接采用论文发表时的 PDF 文件，只是在每篇论文之前加一个简要解释。可以说，这是我现在对当时这些问题与成果的再理解和认识。

这组论文适合高年级本科生、硕士、博士、科技工作者和系统开发人员，或者作为研究生选读教材。读者可以整体下载学习，也可以选择性地下载学习，不一定要逐一阅读。最好的办法是，根据读者对每一篇论文前面的注释的兴趣来确定是否需要阅读这一篇论文。

由于作者水平有限，各种错误难免。比如，这些论文的英文写作水平很一般，在一边研究一边学习写作的过程中完成。另外，因为这些研究问题和方法一直被我思考至今，因此，现在写那些简要解释的时候没有温习相应的论文，也就不能避免有所出入且带有个人偏好性。所以，诚恳欢迎各位读者通过电话或者电子邮件批评指正。

仔细的读者应该已经发现，书名和内容不完全吻合，因为时态区间演算和不确定性推理的矩阵计算不在数据挖掘的研究范畴。选择这两个内容的原因是：没有足够满意的相关成果一起成为一本书，而我认为其研究方法值得推荐给数据挖掘学者。也有请数据挖掘方面的读者帮我推荐给时态推理和不确定性推理学者的想法，我在这里先拜谢了！

张师超
zhangsc@gxnu.edu.cn
2017-05-07 于桂林

目录

算法一：大数据划分挖掘的注释

在“大数据”这个名词出现前，通常使用“大规模数据”和“海量”等来形容数据量的超级大。为了统一名称，本组论文的所有中文注释的描述采用了现在大家熟知的“大数据”这个词，以下不再专门声明。

最初做数据挖掘研究时碰到的第一个难点当然就是数据量大，通常需要几天的时间才能拿到运行结果。所以，我的第一反应就是分而治之！于是，解决这个问题的关键就转变成：如何将划分挖掘的模式融合成能逼近直接挖掘的模式。通过对划分挖掘的模式的仔细分析和理解，我们发现了一种加权融合方法可以逼近直接挖掘的关联模式。它的核心点就是权重的确定：一个关联模式在越多的子集合被挖掘出来，这个关联模式的权值应该越大；一个子集合被挖掘出来的模式中含有权值大的关联模式越多，这个子集合的权值应该越大。这个权重确定的方法得益于美国总统选举办法，每个州的总统选举票数量决定了州的重要性。

请注意，上面提到的“对划分挖掘的模式的理解”是非常重要的，在传统数据挖掘中基本上没有对数据或者数据集合经过理解后再挖掘的算法，通常都是一些应用驱动的挖掘算法。这种理解在某种程度上是对数据集合的理解，导出了一个值得深入探讨的结论：只有对大规模数据划分挖掘并剖析后，才能发现一些新的有用模式，揭示出这些模式的存在形态与内涵。图 1 同时展示了大规模数据的直接挖掘和划分挖掘的过程，后者可以逼近前者的挖掘结果。据我们所知，文献[1]从数学角度证明了这个结论。图 2 展示了划分后在各个数据子集合中发现的局部模式：A 类（蓝色的）、B 类（红色的）和 C 类（绿色的）模式。A 类模式可以采用传统挖掘方法来获取，但是，B 和 C 两类模式是传统算法不能发现的。也就是说，通过这种剖析，我们获得了这样一些新的有趣信息： B 类模式可以用于发现历史数据集合中那些曾经辉煌过的模式，对考证和考古有帮助；C 类模式可以用于发现趋势模式，对市场分析有用。这种理解鼓励了我们展开多源数据（第二篇论文）和动态数据（第三篇论文）挖掘的研究，简单描述如下。

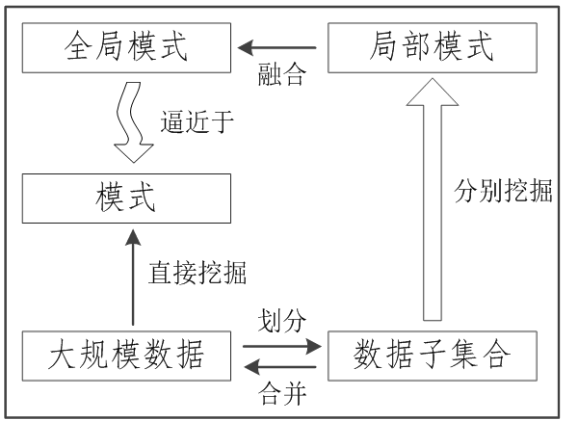

图1 直接挖掘与划分挖掘

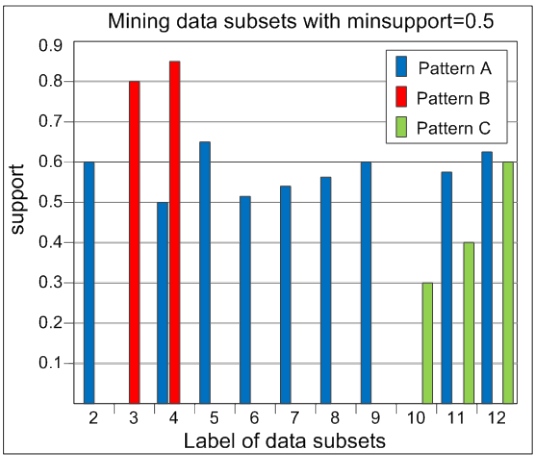

图2 子集合中的模式分布

（1）在多源数据环境下，A、B 和 C 类模式来自于分别挖掘各个数据源，称为

局部模式，它们在局部决策和全局决策中都有重要的应用价值。然而，将数据源集成挖掘的传统方法通常只能发现 A 类型模式中的一部分，因为 B 和 C 两类模式在数据集成后无法识别出来。所以，我们在第二篇论文中提出了局部模式分析（Local Pattern Analysis，LPA）方法，通过基于 LPA 的加权挖掘方法来发现 A、B 和 C 三类模式。

（2）对于一个动态数据集合，我们在第三篇论文中提出了加权挖掘的方法来发现近似的 C 类模式。

参考文献

[1]. C. Xu, Y. Zhang, R. Li, et al. On the Feasibility of Distributed Kernel Regression for Big Data. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11): 3041-3052.

算法二：多源数据挖掘的注释

多源数据挖掘的加权融合方法是与第一篇论文的方法类似，但是，解决拥有多源数据的用户的多层次应用需求问题具有更大的挑战性。

1998 年兴起的多源数据挖掘致力于关联模式发现的方法和技术，符合最初以研究不同类型数据集合中的关联模式挖掘方法为主要目标的数据挖掘领域发展需求。不过，这种多源数据挖掘必须要解决：处理来自多源的数据要面对数据集成所带来的巨大挑战，数据集成导致一些有用模式的消亡，集团公司的多层次挖掘需求，原始数据的私有性保护等问题。这是极为困难的又无法绕开的艰巨任务。针对这个发现，这一篇论文提出了局部模式分析方法，避免了多源的原始数据直接参与运算，实现挖掘复杂性的控制，发现新的全局有用模式，保护原始数据的私有性利益。

基于局部模式分析的多源数据挖掘的具体模式融合过程如下：在任何节点（公司或者子公司）进行模式挖掘，首先将其子节点（如果有）的局部模式作为融合算法的输入，然后，加上从本节点（如果有）的数据中挖掘出的一些模式作为融合算法的输入，最后，通过融合算法有效地综合各种局部模式后产生的模式就是这一层次节点需求的全局模式。这种局部模式分析方法与人类对多源数据的智能处理机制相吻合，可以同时支持全局和局部决策应用，形成了多源多层次模式发现的新机制。与数据集成挖掘方法相比，局部模式分析方法更易于理解与实现，从根本上改变了候选运算数据，减少了存储和计算量，实现了存储和计算难度的控制，保护了原始数据的私有性利益。特别是，局部模式分析能够控制子节点中数据的异质与异构性的传播，完成多源数据挖掘的主体任务。

另外，这一章设计了基于局部模式分析的多源数据中高票模式融合挖掘算法，可以通过调整权重值来扩展该算法，以便发现局部突出模式、例外模式、或者趋势模式（图 2 中的 A、B 和 C 三类模式的具体表现），解决了数据集成后导致这些有用模式消亡的问题。的确，数据集成会导致一些有用模式消亡，下面一个类似的例子可以形象地说明这一点。例如，设 X 和 Y 是争夺网球比赛决赛（3 局 2 胜制）的两名选手，且三局的比分分别是 6：4；0：6；6：4。按照比赛规则，**X 选手取得 2：1 的胜利**（图 2 中的 A 类模式），是冠军。若将这三局比分看成三个数据库的数据，按照数据集成挖掘方法，这三个库中的数据被集成后挖掘到的结果是：**Y 选手取得 14：12 的胜利**，应该获得冠军，但这与赛制规定相左。这是因为将三个库中的数据集成在一起后就不能体现出 2：1 这个三局两胜的局部模式的分布信息，即，数据集成会导致了这类重要信息的消亡。

另外，图 2 中 B 和 C 两类模式通常在数据集成后统计出的支持度很低，也是无法识别出来的。在实际应用中，上面这些隐藏在各个数据源的局部模式是集团公司及其子公司进行决策的最有用的全局信息之一，可以通过扩展这篇论文中的挖掘算法来发现它们，有兴趣的读者可以下载阅读文献[2-4]。

参考文献

[2]. Xindong Wu, Chengqi Zhang and Shichao Zhang. Database Classification for Multi-Database Mining. Information Systems, 30(2005): 71-88.

[3]. Shichao Zhang, Chengqi Zhang and Jeffrey Xu Yu. An Efficient Strategy for Mining Exceptions in Multi-databases. Information Sciences, 165/1-2 (2004): 1-20.

[4]. Chengqi Zhang, Meiling Liu, Wenlong Nie and Shichao Zhang. Identifying Global Exceptional Patterns in Multi-database Mining. IEEE Computational Intelligence Bulletin, Vol. 3 1(2004): 19-24.

算法三：动态数据的增量挖掘的注释

动态是多源数据的重要特征之一，其挖掘结果对实时监测与跟踪、行为预测、模式维护等应用具有重大的实际意义。在传统的动态数据挖掘都是针对数据增加的，但实际应用中数据更新包括了数据增加、删除和修改三类操作。在这一篇论文中，针对增量式动态数据，我们建立了新的竞争机制、权值的自动产生和误差分析。与传统的增量更新算法相比，我们的增量式维护算法是基于加权技术的。通过加权，一些新的、有趋向性的模式能快速脱颖而出，这种模式揭示了市场变化方面的信息，在市场预测等应用中极为有用，但是，传统的数据增加式动态数据挖掘算法不能发现这类模式。

其实，在参考文献中列出了论文[5，6]，方便于感兴趣的读者下载阅读我们的减量式挖掘的模型和方法。这样，通过增量和减量挖掘的有机组合，可以形成数据动态变化的挖掘维护方法。这是实际意义上的一个完备的更新挖掘体系，填补了同时支持增加、删除和修改三类数据更新操作的动态数据挖掘理论与技术的空白。

参考文献

[5]. Shichao Zhang, Jilian Zhang, Chengqi Zhang: EDUA: An Efficient Algorithm for Dynamic Database Mining. Information Sciences, 177(13): 2756-2767 (2007).

[6]. Shichao Zhang, Jilian Zhang, Zhi Jin. A decremental algorithm of frequent itemset maintenance for mining updated databases. Expert Systems with Applications, 36 8(2009): 10890-10895.

算法四：非频繁模式与负关联规则挖掘的注释

自 Agrawal 等在 1993 年提出关联模式挖掘[7]，到 1998 年我开始数据挖掘研究时，这方面的研究成果相当多。读了一些主要论文后，感觉大家基本上是在做 Apriori 算法的改良。除了第一篇论文中想到的划分挖掘方法外，我没有设计出更有意义的频繁模式挖掘算法。经过对实际应用需求和数据的理解，我们发现：传统的关联规则能抓住频繁项集合中相伴性的相互关系，尽管广泛存在且切实有效，但它反映的只是潜在数据项的同现，有着明显的不足。

在现实世界中，完整的关联性应该是双重的，包括同时发生（关联规则）和排斥发生（负关联规则，它表示在某一事务中某些项集合的存在意味着另一些项集合的不存在）这两类关联。互斥性关联模式通常是隐藏在非频繁项集合中的，因此，必须提供一种全新角度的数据相关性分析方法。另一方面，注意到很多现有的数据挖掘算法，比如分类和聚类都可以使用互斥关联分析，所以，互斥性关联方面的成就肯定会对数据挖掘领域带来深远影响。针对以上发现，我们主张充分利用非频繁项集（非频繁模式），从非频繁项集合中挖掘潜在有用的模式：负关联规则，它刻画了项集合中互斥性这一相互关系。与频繁项集合相比，挖掘有用的非频繁项集合更具有挑战性，因为它必须面对一个完全的巨大搜索空间，用于频繁模式的剪枝技术在非频繁项集合挖掘中是无效的。因此，在这一篇论文中我们提出用于挖掘负关联规则的非频繁项集合表示与操作方法、测度理论、挖掘算法、剪枝技术、和评估方法。非频繁模式和频繁模式在实际应用中具有很强的互补性，它们一起形成了数据关联的完整体系，使得一些有意义的非频繁模式能在决策等应用中充分发挥作用。

文献[8]是这一篇论文的会议版本，仅供感兴趣的读者下载阅读。

参考文献

[7]. R. Agrawal, T. Imielinski, A. Swami. Mining association rules between sets of items in massive databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, Washington D.C., 1993: 207 - 216.

[8]. Xindong Wu, Chengqi Zhang and **Shichao Zhang**. Mining Both Positive and Negative Association Rules. In: Proceedings of 19th International Conference on Machine Learning (ICML), Sydney, Australia, 2002: 658-665.

算法五：区间推理的矩阵演算的注释

区间时态推理实际上是一个匹配过程，复杂性很高。通过理解时间区间之间的关系及其演算，我们发现是可以通过矩阵计算来实现。于是，这一篇论文提出了区间关系矩阵，以及演算和判断的方法。这个方法尽管跟算法六的矩阵计算有一些不同，但思想的确是来源于算法六。

文献[9]是算法五的区间演算方法的扩展与深入探讨，仅供感兴趣的读者下载阅读。

参考文献

[9]. **Shichao Zhang** and Chengqi Zhang. Propagating Temporal Relations of Intervals by Matrix. Applied Artificial Intelligence, Vol. 16, 1(2002): 1-27.

算法六：不确定性推理的矩阵计算的注释

在这一篇论文完成之前，我们在不确定性推理方面做了近 10 年的研究。然而，取得的成果一直不满意。所以，我们对不确定性推理的过程展开了分析和再理解，产生了采用矩阵表示规则的可信度的思想。经过反复研究，又定义了这种矩阵的演算和判断方法。实验表明，这种矩阵表示、演算和判断可以很好地实现不确定性推理的匹配过程。我们对这个结果感到满意，就选择出来供读者鉴享。