

大数据

3

于艳华/YU Yanhua, 宋美娜/SONG Meina

(北京邮电大学计算机学院, 北京 100876)

[编者按]数据是与自然资源一样重要的战略资源,大数据技术就是从数量巨大、结构复杂、类型众多的数据中,快速获得有价值信息的能力,它已成为学术界、企业界甚至各国政府关注的热点。本讲座将分3期对大数据进行讨论:第1期介绍了大数据的提出、含义、特点,大数据和云计算的关系以及大数据典型应用;第2期介绍大数据获取、存贮、搜索、分享、分析、可视化等方面的关键技术,并对当前热点技术—可视化进行重点分析;第3期探讨数据流挖掘等实时数据分析技术,介绍大数据中非结构化数据处理和挖掘技术,并给出大数据发展面临的挑战与应用前景。

中图分类号: TN91 文献标志码: A 文章编号: 1009-6868 (2013) 03-0057-06

7 数据挖掘和数据流挖掘

7.1 大数据挖掘技术的简介和分类

大数据技术广义上包括大数据相关的获取、存储、处理、挖掘等技术,但就美国政府2012年提出的“大数据研究与发展计划”而言,它主要指的是面向大数据的数据挖掘、机器学习技术。此期重点介绍大数据中的数据挖掘技术,重点是数据流挖掘技术。

数据挖掘技术是一个涉及数据库、机器学习、统计学、神经网络、高性能计算和数据可视化的多学科领域,是计算机模仿人类学习机理和方法,利用数据自动获取知识的一种技术。数据挖掘出现于20世纪80年代末,在过去的20年中得到了广泛的研究和快速的发展,表现在出现了大量的算法,并可以处理各种类型数据。然而随着大数据时代的来临,数据挖掘技术迎来了空前广泛的应用机会,也面临新的挑战。大数据是伴随智能终端的普及和互联网上微博、社交网络等业务的广泛应用而出现

的,因此面向大数据的数据挖掘的应用首推Google、Amazon、Yahoo、阿里巴巴等互联网公司,比如2009年甲型H1N1流感爆发时,Google利用海量的用户搜索词及其组合,比美国国家疾控中心更及时更准确地报告了疫情;Amazon公司首先提出并应用协同过滤技术进行书籍推荐,其应用效果超过了之前被誉为“公司皇冠之上宝石”的书评团队,开启了电子商务应用中商品推荐的先河。基于互联网上海量语言材料应用机器学习技术的Google语言翻译系统,则是目前为止最为成功的计算机自动翻译系统。面向大数据的数据挖掘技术的一个挑战是:大数据时代我们能得到现象相关的所有数据,即统计学上所说的总体,而不再是传统的统计学和数据挖掘中一个容量有限的样本或容量有限的训练集。另外一个挑战是所得到的数据不是绝对精确的,只要在保证速度的前提下近似地反映宏观和整体情况^[12],这一挑战要求数据挖掘要能处理非结构化数据和含噪音的数据,而挖掘结果的正确性则只要保证在期望的区间内。目前来看,应对这两个挑战的主要技术之一就是

数据流的挖掘。

数据挖掘技术主要分为如下几个分支:分类、聚类、关联规则挖掘、序列模式挖掘、异常点挖掘、时间序列分析预测等。在大数据的相关挖掘应用中,虽然处理的数据形式更丰富,但就学习方法来看并没有根本差别,因为全部是基于数字化后信息的学习。

7.2 概念漂移

“概念漂移”是Schlimmer等人于1986年首次提出的^[13]。大部分的数据挖掘技术都有一个假设前提:样本是随机获取的,并且服从同一稳定的分布。然而在大数据场景下,数据源源不断地到来,样本具有不稳定和不确定性。例如,顾客的兴趣随着时间很有可能发生变化;用户上网的浏览习惯也会随着时间的推移而发生明显地改变。因此大数据场景中不可避免的,一定要考虑概念漂移问题。如图8,样本的统计特性在某一时刻开始发生变化,我们认为此时发生了“概念漂移”。

从样本是否服从相同分布的维度,可以将数据流划分为2类:稳定

收稿日期: 2013-03-21

网络出版时间: 2013-05-07

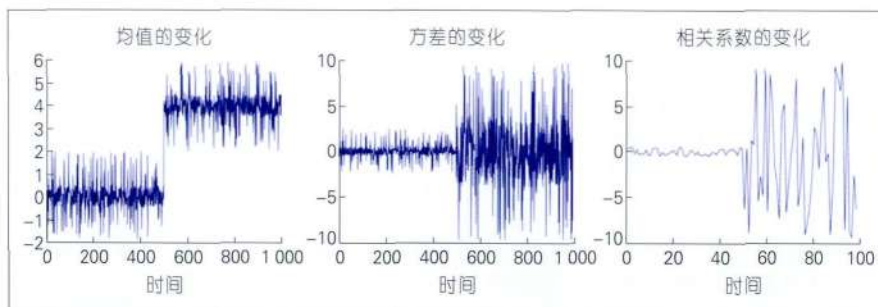
数据流,样本服从同一分布;动态数据流,随着时间推移,样本服从不同分布,只有动态数据流中才存在“概念漂移”现象。概念漂移又可以分为:突变式和渐变式,对这两种漂移的处理方式和难度通常并不相同,在设计漂移算法时,应该分别进行考虑。如图9所示,在 t_0 时刻之前,数据样本服从同一分布A,而在 t_0 和 t_1 之间,数据流发生概念漂移,在 t_1 时刻之后,数据重新趋于稳定,并服从同一分布B。

当概念漂移发生之后,最直接的结果就是从之前样本中学习获得的概念模型,已经不再适用,必须尽快更新。现有概念漂移检测的方法,可以分为3类:模型性能监测法、概念聚类法、样本分布监测法。

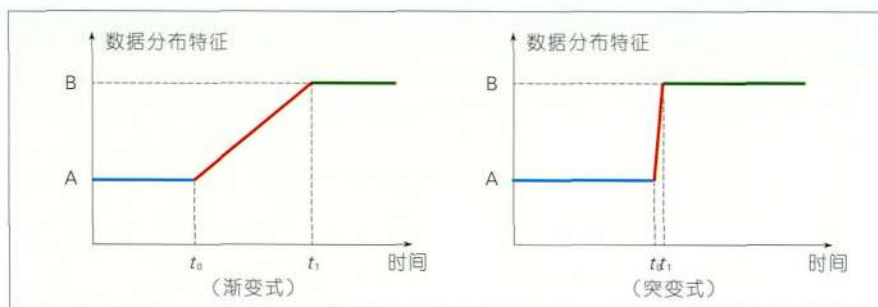
(1)模型性能监测法。以分类挖掘为例,首先需要对分类器的性能进行跟踪监测,当使用新采集的训练集,对现有分类器进行更新之后,如果分类器在测试集上表现出的性能明显下降,我们则认为发生了概念漂移。Windmer 和 Kubat 提出的 FLORA 系列算法^[14]、Last 提出的 OLIN 算法^[15]等都是属于这一类。模型性能监测是十分常用的方法,但当数据流中存在类别不平衡或者进行半监督学习时,此方法将不再适用。

(2)概念聚类法。Katakis 在 2010 年首次提出这一方法^[16],基本思路是将数据流划分为数据块,并且再将其映射为“概念向量”,对多个概念向量进行聚类,每一个聚类代表一个概念。当一个新的数据块到来时,计算其对应的概念向量与各个聚类中心之间的距离,并以此判断是否发生了漂移。这一方法可以解决概念漂移领域的一个重要问题:重复概念的检测。概念聚类法局限的地方在于:假设每次划分的数据块内所有数据都属于同一概念。

(3)样本分布监测法。针对样本集,提取其中的统计特性:特征值分布等,以这些参数的变化来判断是否



▲图8 概念漂移



▲图9 概念漂移:渐变式和突变式

发生概念漂移。2006–2011年间, Alippi^[17-18]、Peter^[19]、Kuncheva^[20]等人都是基于此原理提出了检测概念漂移的具体策略。

7.3 聚类

Han Jiawei 教授在《Data Mining: Concept and Techniques》中,对聚类有一个简短的定义:将物理或抽象对象的集合分成相似的对象类的过程称为聚类。更形式化的一个描述方法是:聚类分析就是按照某种相似性度量方法对对象进行分组,使得各组内的相似度高,而组间的相似度低。俗语“物以类聚,人以群分”可以说是聚类作用的一个生动说明。

聚类挖掘已广泛用于各种应用领域的模式识别以及离群点检测中。市场分析人员可以在没有任何先验知识的情况下,应用聚类方法基于购买模式数据库发现不同的顾客群;网络数据分析人员针对web文档数据或网络访问日志数据对访问的网页进行聚类,以发现对不同网页信息感兴趣的人群,来支持精准营销或分析社会学上原因。应用聚类还可

以发现异常点,即那些无法归入任何簇的点,离群点检测广泛应用于信用卡欺诈检测和监控电子商务中的犯罪活动。聚类分析还可以作为研究数据分布的功能以及作为其他算法的预处理步骤。

从1967年研究人员提出第一种聚类算法开始,目前为止已经有多种可用的聚类算法。但是没有任何一种是普遍适用的,因为不同问题中数据的维度高低不同、各维数据特性不同、数据分布情况不同、数据规模不同,而随着大数据时代数据流的出现,对聚类算法更提出了内存限制、处理时间限制等挑战。但这些算法可以按照聚类依据不同进行分类,首先总体分为2大类:基于样本的聚类、基于变量的聚类。其中,基于样本的聚类人们研究的比较多,前面的聚类举例也全部是针对基于样本的;基于变量的聚类顾名思义就是对变量(即维度或属性)进行分组,它和数据分析中的因子分析及主成分分析(PCA)比较像;但聚类分析并不会对变量进行合并,只是用层次式等方法对变量的远近亲疏程度进行判别。

在某些领域,基于变量聚类非常有用,比如传感器网络、社会网络、电力供应、股票市场上,比如通过聚类分析我们可以发现各支股票之间的关系,而通过流数据聚类则可以发现这种关系的变化情况。

基于样本的聚类是目前为止研究的最多,这些算法又可以分为:基于划分的聚类、基于层次的聚类、基于网格的聚类、基于密度的聚类、基于模型的聚类。对流数据的聚类也是在这些聚类算法的基础上发展而来的,因此,接下来简要介绍下这几类聚类算法及其特点。

7.3.1 基于划分的聚类

经典的聚类算法 k -means 就是基于划分的,这种算法之所以应用广泛是因为其简单快速。但该算法需要人为设定一个代表聚类个数的参变量 k ,如何正确设置这个值是个难题。另外, k -means 算法的理论基础是找到 k 个点(所谓中心点 centroid)使得相应簇中的点到这 k 个点的距离平方和最小。由此可见,采用这种理论所找到的簇是球形的,而且这种方法对噪声和孤立点敏感。而 k -中心点法则是克服了这个问题的另一种基于划分的聚类算法。为了处理大规模数据集,人们在这些算法基础上进行了改进,提出一些新的算法如最大期望算法(EM)、基于随机选择的聚类算法(CLARANS)等。

对数据流聚类时,因为流数据不断到达,所以无法在数据完全到达后进行聚类,部分数据上的聚类结果也很可能不再适用后面到达的数据,因此必须进行增量式聚类。而且,为了及时对后面很快到达的数据进行处理,每次的聚类操作必须在指定时间内完成,同时内存也要不断腾出来配合下一次聚类操作。当然,聚类结果可能达不到理论上的完美效果,但是要有尽可能好的效果,最好这个结果和理想结果差多少有一个理论上的范围。这些问题其实是所有流数据

挖掘和静态数据的区别所在:要在有限内存有限时间内给出一个准确性有一定保证的挖掘结果,

Farnstrom 等人提出的一趟 k -mean 算法是适应流数据挖掘的 k -means 算法,它只对数据进行一趟扫描,当然历史结果的保存需要采用一种叫做聚类特征的概要数据。Domingos 和 Hulten 在此基础上提出的快速 K 均值算法(VFKM)则对每次增量聚类时需要的样本个数给出了理论上计算方法,其采用的理论基础是 Hoeffding 不等式,这个不等式和契比雪夫不等式性质类似,都是对于一个分布特性未知的随机变量,已知很少量的统计参数,可以在任意置信度之下计算出相应的置信区间。而 Guha 等人则提出了数据流聚类的 k -中心点算法,并给出所需的样本个数及所需时间和空间的理论计算结果。

7.3.2 基于层次的聚类

层次聚类也是一种常用聚类方法。它不再是只给出 k 个聚类而成的簇,而是给出多层的树状聚类结果。层次聚类又可分为凝聚和分裂两类,分别采用自底向上和自顶向下两种方法。BIRCH 算法则综合了这两种方法。

Aggarwal、J. Han 等人提出的 CluStream 算法则是 BIRCH 算法在数据流挖掘上的扩展。该算法的特征之一是:提出了倾斜时间窗口的概念,依据较近的数据比历史数据更重要的理念,最近的时间变化以较细的时间粒度刻画,而离现在较远的数据则采用较粗的时间粒度。该算法的另一个重要特点是,整个流聚类分为在线和离线两部分。在线部分增量式进行数据处理,获得摘要信息微簇(micro-cluster),离线部分宏簇(macro-cluster)通过对在线部分的结果进行再处理获得层次的聚类结果。

7.3.3 基于网格和密度的聚类

基于密度的聚类不再按之前两

种聚类采用的距离的远近作为分划的依据,而是按照单位空间范围内点的个数即密度来划分空间,只要某一范围内密度大于某一指定参变量,则认为是同一簇。基于密度的聚类算法(DBSCAN)、通过对对象排序识别聚类结构算法(OPTICS)等是经典基于密度聚类算法。

基于网格的聚类是面向时空相关问题。它采用一个多分辨率的网格数据结构,这些网格把空间量化为有限数目的单元,所有聚类操作都在这些网格上进行。这些方法的主要优点是处理速度快,独立于数据对象数目,只与每一维上的单元数目相关。经典算法是信息网格算法(STING)、WaveCluster,而 Quest 上聚类(CLIQUE)则综合了密度和网格两种方法。

在流数据聚类中,分形聚类则是一种基于网格的聚类,它将具有相同分形维的具有高自相似性的点分为一类。

7.3.4 基于模型的聚类

基于模型的聚类其实是把回归拟合应用在聚类中,它为每一簇拟合一个模型,根据拟合模型的方法不同又分为统计学方法和神经网络方法,属于前者的有简单增量概念聚类算法(COBWEB)方法,属于后者的有学习矢量量化网络(LVQM)、自组织映射(SOM)等方法。

7.4 数据挖掘中的分类

数据挖掘中的分类指的是:首先根据已知类别的一些样本进行学习,得到一个分类的规则或者说是模型,然后利用学习得到的模型对另外一些类别未知其他属性值已知的样本进行类别的判断或者预测。可以看出,分类和聚类的不同之处在于:分类学习时,样本类别是已知的;而聚类学习时,样本类别甚至类别数目是未知的。因此前者是有监督的学习,后者则是一种无监督的学习。分类

学习的一个经典的例子是对银行现有的顾客信用信息进行学习,建立信用良好或欺诈客户的判断模型,当一个新的顾客申请银行借贷时,利用学习模型进行判断,给出新客户良好或是欺诈客户的可能性,从而提高银行业务决策的科学性。

典型的分类方法有很多,主要包括基于决策树(DT)的分类、基于贝叶斯(Bayesian)分类、基于神经网络的分类等。决策树分类是基于信息论中的信息熵的概念,学习结果是一个由各个属性及其取值形成的代表判断流程的树状结构,称为决策树。典型的算法包括ID3、C4.5等。适用于大规模数据集决策树构造的算法则有Quest上的有监督学习(SLIQ)和可伸缩并行决策树(SPRINT)等。贝叶斯分类算法基于统计学中的贝叶斯后验概率定理,并应用各属性间类条件独立的朴素假定,方法简单,可伸缩性好,很多实验表明其分类效果与复杂的决策树和神经网络相媲美。

传统的分类方法多是非增量式的,即当全部训练样本准备好之后,对样本集进行多次扫描,获得一个分类器,例如工业界广泛应用的分类算法C4.5和CART;而数据流场景下,由于数据源源不断地到来并且数据量巨大,完全将数据存储下来再进行处理,是无法实现的,这就要求分类算法必须是增量式的,即训练样本集不能一次性全部获取的情况下,先利用已经获得的样本集来建立分类器,再用新到达的样本来修正分类器。

快速决策树算法(VFDT)是由Domingos、Hulten等人在2000年提出的^[20],主要用于解决稳定数据流的分类问题,性能渐近逼近传统的C4.5算法,其基本思路为:利用Hoeffding不等式来保证选取的分裂属性的可信程度,并且不断地将叶子节点替换为中间节点(决策节点),最终生成一棵决策树。其中每个叶节点都保存着样本属性值的统计信息,这些信息将用于选取分裂属性。当一个新样本

到来后,它将沿着决策树从根节点向叶节点去遍历,它在树的每个中间节点都进行属性值判断,并进入不同的分支,最终到达叶节点,并更新叶节点上的统计信息。每隔一段时间重新评估每个叶节点,选取满足Hoeffding不等式的属性,进行分裂。

现在通过一个简单的实例,来说明VFDT算法的基本过程。如图10所示,假设从 t_0 时刻开始进行挖掘,样本源源不断地到来,此时节点1是叶节点(根节点),样本到达节点1之后,更新其中的属性值统计信息,并判断是否有属性满足Hoeffding不等式;假设在 t_1 时刻,一个样本到达后,节点1内某一属性满足Hoeffding不等式,则按照此属性进行分裂,产生节点2和节点3,节点1由叶节点变为中间节点;此时, t_0 到 t_1 之间所有到达样本的统计信息,都被舍弃;从 t_1 时刻起,所有新到达的样本数据,根据节点1中的属性分裂条件,到叶节点(达节点2或者节点3),并更新叶节点中的统计数据,同时判断是否有属性满足Hoeffding不等式,若有则继续进行分裂生长。从上述过程可以看出,决策树每次进行生长时,都会单独占用并消耗一部分数据:节点1分裂时,消耗了 t_0 到 t_1 之间所有到达节点1的样本,这些样本将不再对此后决策树的生长产生任何影响;当节点2分裂时,消耗了 t_1 到 t_2 之间所有到达节点2的样本,这些样本将不再对此后决策树的生长产生任何影响。

基于VFDT算法,Hulten、Domingos等人于2001年提出可以解决概念漂移问题的概念自适应快速决策树算法(CVFDT)。此后近十多年时间里,

针对VFDT算法拓展和应用的层出不穷,CVFDT算法都取得了不错的性能测试效果。然而在2012年Rutkowski等人在TKDE上发表一篇文章指出,VFDT算法中使用的Hoeffding界不符合数据流的应用场景,应该改为McDiarmid's界^[22]。这一点感兴趣的读者可以自己查阅,但不可否认的是在各式各样的测试数据集上,VFDT确实显示出令人满意的测试性能。

此外,数据流中经典的分类算法还有:基于模糊信息网络的2002年Last提出的OLIN算法等。特别要说明的是,近几年在数据流分类挖掘中,基于单分类器的集合分类器方法得到了较广泛的研究和应用。

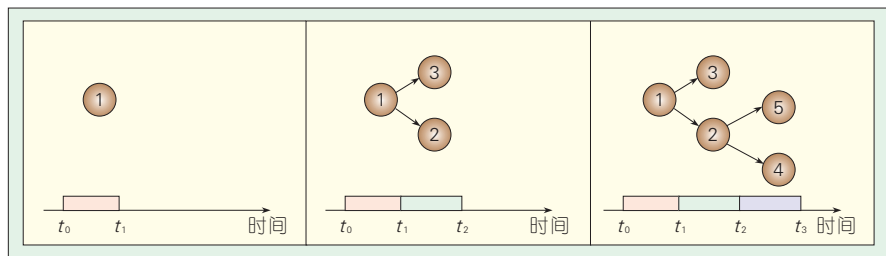
7.5 频繁模式挖掘

7.5.1 关联规则挖掘算法

关联规则挖掘算法的基本概念包括两个方面的内容:项以及项集,其中项是基本单元,用来表示实际环境中的单个具体事物,例如在超市购买的物品;项集是由一个或者多个项组成的集合,表示的是具体的一次事务,例如顾客的一次购买行为,在项集内部,项与项之间不存在次序关系。而所谓的关联规则是形如 $X \rightarrow Y$ 的蕴涵表达式,其中 X 和 Y 是不相交的项集,即 $X \cap Y = \phi$ 。通常的关联规则算法主要分为两个步骤:

(1)产生频繁项集。其目标是发现满足最小支持度阈值的所有项集,并将这些项集称为频繁项集。

(2)产生关联规则。分解频繁项集,获取满足最小置信度的规则集,并将这些规则称为关联规则。



▲图10 VFDT算法

其中,支持度表示给定数据集的频繁程度,而置信度是指在包含的事务中出现的频繁程度。

关联规则算法是由 R.Agrawal 首次提出的,称为 Apriori 算法。它采用“支持度—置信度”的框架产生关联规则集,其影响深远,后续许多算法都是基于其思想提出的,并统称为类 Apriori 算法。该类算法首先是利用 k -频繁项集,计算得到对应的 $(k+1)$ -候选项集;其次利用先验定理(频繁项集的子集一定是频繁项集)裁剪非频繁项集;最后使用支持度裁剪机制获取 $(k+1)$ -频繁项集。之后重复上述迭代过程,直到无法产生新的频繁候选项集为止。其算法的缺点是产生过多的候选项集,并且多次扫描数据库。

另一个有影响深远的算法是 FP-growth 算法,针对 Apriori 算法多次扫描数据库的缺点,FP-growth 算法设计了一种 FP-Tree 的数据结构体,通过读取一次数据库将其所有的数据压缩到一棵 FP-Tree 上,并通过循环产生前缀序列的 FP-Tree,获取对应的频繁项集。该算法的优点在于利用 FP-Tree 结构压缩原始数据集,缩小搜索范围,快速产生频繁项集。

通过多年的发展,目前关联规则算法已经定义了许多新类型的模式,如模糊关联规则、稀有关联规则、基于权重的关联规则等。由于关联规则算法的日趋成熟,其相应的研究热点已经从如何产生关联规则逐渐转变为如何产生有效的关联规则,例如目前有效规则的一个研究热点是如何挖掘高“效用”的关联规则^[23]。

7.5.2 频繁序列模式挖掘算法

频繁序列模式挖掘算法是由 Agrawal 和 Srikant 首次提出的,并且随着其被广泛应用在分析用户的购物习惯、异常行为检测以及网络入侵检测等应用场景中,序列模式挖掘算法的研究取得了迅猛发展。从宏观上讲,序列模式的组成包括3方面的内

容:序列、事件(事务或者项集)以及项,它们三者之间的关系是序列是由一个或者多个事件组成的,而事件是由一个或者多个项组成的;在组成序列的事件中,事件与事件之间存在着先后时间关系,而在组成事件的项中,项与项之间不存在先后时间关系。

频繁序列模式依据产生序列模式的方法不同可以分为两种:一种可以被称为类 Apriori 算法,其基于“候选—测试”的思想,利用前一步产生的 k -频繁序列模式,产生 $(k+1)$ -频繁序列模式候选集,并利用支持度测试的裁剪机制,从而获取最终的 $(k+1)$ -频繁序列模式集。其具有代表性的算法包括:AprioriAll 以及 SPADE^[24]算法,其中图 11 展现了使用 SPADE 算法产生新的候选序列的过程。

如图 11 所示,SPADE 算法使用树形结构,利用上层的 2-频繁序列模式 a_1-b_1 以及 a_1-d_1 产生 3-频繁序列模式 $a_1-b_1-d_1$ 。类 Apriori 算法的优点是可以挖掘出在限制条件下所有的频繁序列模式集,其缺点是有些类 Apriori 算法会在产生频繁序列模式集的时候,多次扫描数据库,增加算法的 I/O 操作;其次在产生频繁序列模式的时候,会产生大量的无用候选序列,增加算法的计算时间,降低算法的挖掘效率。

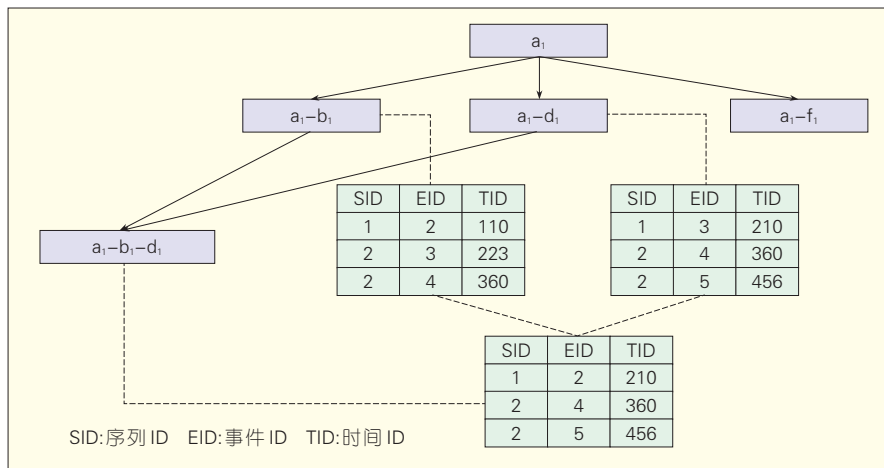
另一类算法是采用“投影”技术,依据不同的前缀序列对原始数据集

进行划分,并通过不断更新前缀序列以及划分数据集的操作,最终获取完整的频繁序列模式集,其具有代表性的算法是 PrefixSpan^[25]。图 12 显示了利用“投影”技术,获取的原始数据集中所有 1-前缀序列所对应的投影数据库:

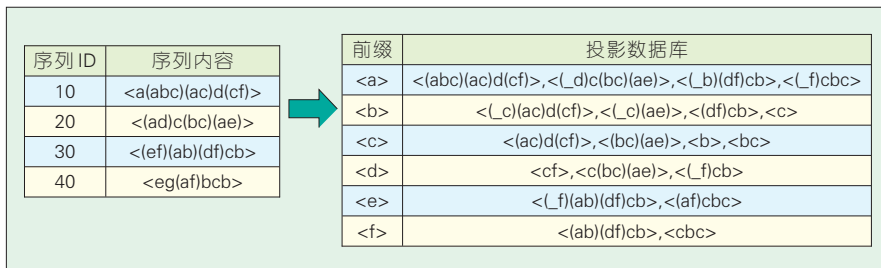
在图 12 中显示了利用“投影”技术,获取原始数据集对应的所有 1-前缀序列的投影数据库。其算法的优点在于利用“投影”技术可以将原始数据集的规模不断缩小,以缩小算法的搜索范围,同时由于各个前缀的投影数据库是相互独立的,所以可以并行地挖掘对应的各个投影数据库,提高算法的挖掘效率;该算法的缺点是如果前缀序列在原始序列集中分布均匀,即对应的投影数据库变小趋势缓慢,则无法缩小算法的搜索空间。根据算法挖掘结果的不同,可以将序列模式算法分为:全集频繁序列模式挖掘算法、闭合频繁序列模式挖掘算法以及最长频繁序列模式挖掘算法等。

7.5.3 基于数据流的频繁序列模式挖掘算法

由于数据流具有无限性以及动态性的特点,因此传统的频繁序列模式挖掘算法已经无法适用于数据流对象,如何在数据流中获取频繁序列模式已经成为了序列模式挖掘算法



▲ 图 11 SPADE 算法新频繁序列生成过程



▲图12 原始数据集转变为投影数据库

中的一个研究热点,由于其尚处在一个发展阶段,大部分的算法还是在原有的数据流基本算法的基础上,结合序列模式挖掘算法设计完成的。根据使用不同基本算法,数据流挖掘算法大致可以分为3类,第1类是利用给定的界限值,挖掘近似的频繁序列模式集;第2类是设计一种新的滑动时间窗口,基于批处理的思想,挖掘频繁序列模式集;第3类是设计一种新的数据结构,例如FP-Growth中的FP-Tree结构体,保存对应的压缩信息,结合滑动时间窗口,挖掘频繁序列模式集。根据数据流动态变化的性质,又可以将数据流挖掘算法分为两类,一类是针对分布固定不变的数据流对象,挖掘近似完备的频繁序列模式集,另一类是针对动态分布变化的数据流对象,检测数据流中出现的“概念漂移”的现象,解决模型失效的问题。

8 结束语

物联网兴起,互联网高速发展,各种信息普遍数字化,PB级数据广泛出现,云计算和云存储技术都在改变人们使用计算机使用信息服务的方式,企业依托海量数据学习来解决以往无法解决问题,互联网企业则利用数据挖掘技术获得高额利润和社会影响力,这些都意味着大数据时代的来临。大数据的获取和应用对企业来讲,意味着经济效益,Google、Yahoo、阿里巴巴等是大数据应用获益的典型代表;对科技界来讲,意味着新的科学研究方法甚至是新的科研范式;而大数据对政府而言则是与

人力资源、自然资源一样重要的国家战略资源。但是,在大数据的研究和应用中,存在着很多问题和挑战,包括:(1)传统关系数据模型无法高效处理非结构化和半结构化数据,以MapReduce和Hadoop为代表的非关系数据分析技术在应用性能等方面仍存在很多问题,尚没有一个像当年Codd所提出的关系数据库那样的理论来统一解决非结构化处理问题。(2)适合不同行业的大数据挖掘分析工具和开发环境。不同行业需要不同的大数据分析工具,当前跨领域跨行业数据共享仍存在很多壁垒。(3)数据隐私保护。大数据以数据的共享为基础,但如何同时保护用户的隐私则是需要解决的问题。相信随着大数据技术问题逐步解决,大数据应用必将给我们社会和生活带来更多的正能量。

参考文献

- [12] MAYER-SCHONBERGER V, CUKIER K. 大数据时代:生活、工作与思维的大变革[M]. 盛杨燕,周涛,译. 杭州:浙江人民出版社,2012.
- [13] Schlimmer J C, Granger R H Jr. Incremental Learning from Noisy Data[J]. Machine Learning, 1986, 1(3): 317-354.
- [14] Gerhard W, Kubat M. Effective Learning in Dynamic Environments by Explicit Context Tracking[C]//Proceedings of the European Conference on Machine Learning (ECML'93), Apr 5-7, 1993, Vienna, Austria. Berlin, Germany: Springer, 1993.
- [15] Last M. Online Classification of Nonstationary Data Streams[J]. Intelligent Data Analysis, 2002, 6(2): 129-147.
- [16] Katakis I, Tsoumakas G, VLAHAVAS L. Tracking Recurring Contexts Using Ensemble Classifiers: An Application to Email Filtering[J]. Knowledge and Information Systems, 2010, 22(3): 371-391.
- [17] Alippi C, Roveri M. Just-in-time Adaptive Classifiers—Part II: Designing the Classifier [J]. IEEE Transactions on Neural Networks, 2008, 19(12): 2053-2064.

- [18] Alippi C, Boracchi G, Roveri M. An Effective Just-in-Time Adaptive Classifier for Gradual Concept Drifts[C]// Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'11), Jun 31-Aug 5, 2011, San Jose, CA, USA. Piscataway, NJ, USA: IEEE, 2011: 1675-1682.
- [19] Vorburger P, Bernstein A. Entropy-Based Concept Shift Detection[C]// Proceedings of the 6th IEEE International Conference on Data Mining (ICDM'06), Dec 18-22, 2007, Hong Kong, China. Los Alamitos, CA, USA: IEEE Computer Society, 2006: 1113-1118.
- [20] Kuncheva L I. Change Detection in Streaming Multivariate Data Using Likelihood Detectors[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(5): 1175-1180.
- [21] Domingos P, Hulten G. Mining High-Speed Data Streams[C]//Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00), Aug 20-23, 2000, Boston, MA, USA. New York, NY, USA: ACM, 2000: 71-80.
- [22] Rutkowski L, Pietruczuk L, DUDA P. et al. Decision Trees for Mining Data Streams Based on the McDiarmid's Bound[J]. IEEE Transactions on Knowledge and Data Engineering, To be published.
- [23] Tseng V S, WU C W, Shie B E, et al. UPGrowth: An Efficient Algorithm for High Utility Itemset Mining[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), Jul 25-28, 2010, Washington, DC, USA. New York, NY, USA: ACM, 2010: 253-262.
- [24] Zaki M J. SPADE: An Efficient Algorithm for Mining Frequent Sequences[J]. Machine Learning, 2001, 42(1/2): 31-60.
- [25] Pei J, Han J W, MORTAZAVI-ASL B, et al. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth[C]//Proceedings of the 17th International Conference on Data Engineering (ICDE'01), Apr 2-6, 2001, Heidelberg, Germany. Piscataway, NJ, USA: IEEE, 2001: 215-224.

作者简介



于艳华, 北京邮电大学计算机学院副教授; 主要研究方向为网络管理与优化、数据挖掘等; 已发表论文10余篇, 申请专利10余项。



宋美娜, 北京邮电大学计算机学院教授; 主要研究方向为分布式系统、服务计算、数据工程等; 已发表论文50余篇, 申请专利20余项。