数据流分类挖掘中的概念变化研究

韩法旺 刘耀宗
(南京森林警察学院信息系 南京 210023)

摘 要 数据流分类挖掘首先要面对概念变化问题。介绍了数据流分类中的概念变化的定义与类型,研究了概念变化的意义及应用,对目前数据流中处理概念变化的方法进行了综述。真实数据流常常含有大量的噪声,因此需要理解噪声与概念变化的区别。针对周期性数据流中概念重现现象,当"历史概念"重现时,利用特定的模型对数据流进行概念预测,可以减少模型更新的代价。

关键词 数据流分类,概念变化,概念重现,噪声
中图法分类号 TP181 **文献标识码** A

Study on Concept Change in Data Streams Classification

HAN Fa-wang LIU Yao-zong
(Department of Information, Nanjing Forest Police College, Nanjing 210023, China)

Abstract Data stream classification must face the concept of change. This paper introduced the definition and types of conceptual changes in the data stream classification, the meaning and application of conceptual changes, and the methods of conceptual changes in the data stream. Real data stream often contains a lot of noise, and needs to understand the difference between noise and the concept of change. To reproduce the phenomenon for periodic data stream concept, when "the concept of history" reproduces, the concept of prediction using a specific model of the data stream can reduce the model update price.

Keywords Data streams classification, Concept change, Recurrent concepts, Noise

1 引言

数据流分类技术已经成为数据挖掘的研究热点[1],数据流分类和传统的分类挖掘有着很大的不同,数据流中分类挖掘的研究是当前数据流挖掘领域的重要内容之一。数据流分类必须要首先解决概念变化(Concept Change)的问题。概念变化通常是指隐含内容的改变会或多或少从根本上导致目标概念的改变。由于数据流的动态性,其所隐含的映射关系会发生变化,而这种蕴含在训练数据中的映射关系的变化可看作是发生了概念变化[2]。

如何解决数据流分类中的概念变化问题,也是数据流分类挖掘研究的难点。数据流分类算法大致可分为两类[1]:增量学习(Incremental Learning)和集成学习(Ensemble Learning)。这两种方法对概念变化处理仍然有很大的局限性,一般笼统地把数据流中的概念变化看成是样本的联合概率分布发生变化,并没有深入地探讨概念变化的起因和相应的解决办法。

目前对数据流概念变化的研究大多数是数据流分类研究中的附带解决的问题之一,还没有针对数据流挖掘中概念变化现象做过系统性归纳。处理概念漂移的数据流主要有以下要求[3]:1) 准确性。数据流分类算法中广泛采用的遗忘机制很难与概念变化的检测机制做到精度与效率的统一;2) 有效性。现有算法存在对轻微的概念变化过于敏感、判断漂移的计算代价较高的缺点,而对于概念变化的速率与数据流分类之间的效率关系缺少深入研究。3)概念变化与噪音数据的关系。真实数据流往往并存大量的噪音数据,这给数据流挖掘带来诸多困难,而数据流分类器在处理概念变化时必须处理好噪音数据的影响。

总的来讲,目前正在深入研究概念变化问题,但还缺少对概念变化的形式化的描述,没有形成公认统一的看法。本文针对数据流分类挖掘中的概念变化研究作了阶段性的归纳;深入分析了数据流中的概念变化现象与机理;对数据流中概念变化出现的新问题进行了探讨,并提出了相对应的策略。

2 数据流中概念变化的含义

2.1 数据流的概念变化含义

研究概念变化先要理解概念的含义,概念学习是机器学习领域的重要研究内容,许多机器学习问题都涉及到从特殊训练样本中得到一般概念,待学习的概念或函数称为目标概念(Target Concept)[4]。

定义 1(概念学习[4]) 给定一个总体样本空间 X,X 中的每个样本表示为特征属性的集合,待学习的概念为 c,c 是定义在 X 上的概念函数,从 X 中抽取训练样本集合 T(T 中每个样本的目标概念值 $c(x)$已知),概念学习的任务就是在

本文受中央高校基本科研业务费专项资金项目(LGYB201412)资助。
韩法旺(1972—),男,硕士,讲师,主要研究方向为计算机网络;刘耀宗(1975—),男,博士,主要研究方向为数据流挖掘与管理,E-mail:new025@126.com(通信作者)。

只知道训练样本的目标概念 c 的情况下，在总体样本空间 X 上确定与目标概念 c 相同的假设 h，满足 $\forall x \in X, h(x) = c(x)$。

由于数据流是无限且不断变化的，因此不可能给定一个完整的总体样本空间去分析研究，而一般是从“可能引起概念变化发生的原因”或“概念变化发生后可能引起的后果”两点来分析预测是否有概念变化发生。

定义 2(概念变化[5]) 数据流上的概念变化描述的是在数据流 S 上的一个变化的目标概念。设 S_i 是在时刻 i 处接收的流数据，$FO_i(x)$是最佳分类模型，设 $FO_{i-1}(x)$是在前时刻 $i-1$ 处前一个最佳分类模型。如果 $FO_{i-1}(x)$与 $FO_i(x)$是不一致的，则称从时刻 $i-1$ 到时刻 i 存在着概念变化。

概念变化有 3 种模式[6]：概念漂移(Concept Drift)、概念转移(Concept Shift)和采样变化(sampling change)，概念漂移是渐变的(Gradual)，而概念转移是突变的(Abrupt)，第三种方式采样变化则是数据分布的变化，数据分布的变化导致当前模型需要修正，也被称作虚拟概念变化。虚拟概念变化表示样本的变化，而实际概念变化表示内容主题的变化。虚拟概念变化和实际概念变化可能会同时发生。无论发生的是哪种概念变化，即使概念保持不变也要修正当前的模型。真实的数据流环境中，真实概念变化和虚拟概念变化往往是同时并存的[7]，这使得概念变化对分类器的影响难以度量。但是如果样本的分布发生了变化，分类器在此样本上的误差率将会增加，即发生何种概念变化并不重要，我们只需要检测分类器在当前分布上的误差率就可以判断两个分布是否一致。

数据流的概念变化检测可以通过概念学习获取 t 时刻的概念 $c(t)$，当新数据在 $t+\Delta t$ 时刻流入时，概念变化检测方法获取此时刻训练集中概念 $c(t+\Delta t)$，当 $c(t) \neq c(t+\Delta t)$时，则概念变化就在 $t+\Delta t$ 时刻发生。但往往直接比较概念的异同是非常困难的，学习一个不断变化的概念更不可行，况且概念学习的前提是“目标概念”是已知的。而在实际应用中，数据流中存在的多概念变化通常是指隐含内容(hidden context)[7]的改变会或多或少从根本上导致目标概念的改变。而在实际应用中，由于噪声数据(Noise)的大量存在，数据流中究竟存在多少种概念、存在何种概念都是未知的[8]。

图 1 是数据流中噪声与概念变化的示意图。C1，C2 表示数据流中不同的概念。

Noise 表示数据流中的噪声数据，Noise 数据的特点是数据无规律，不同于概念变化；Abrupt 是突变类型的概念变化；Gradual 是渐变类型的概念变化；Recurring 表示周期性数据流的概念重现。

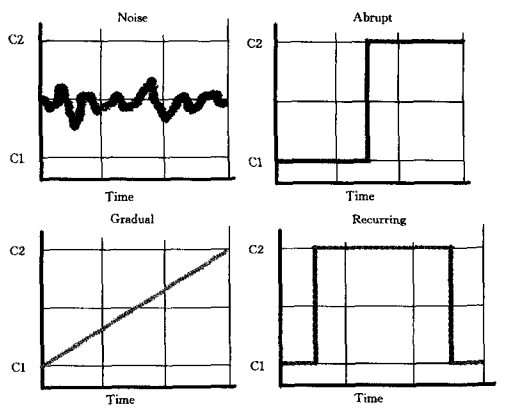

图 1 数据流中概念变化的类型

2.2 噪声与概念变化

真实的数据流都含有噪声，数据流中概念发现的实质就是从含有噪声的数据流中提取被研究对象的数学模型，而概念变化的实质就是被研究对象的数学模型的变化及其变化程度。噪音的存在可能导致挖掘结果精确度低，甚至产生错误的结果。而实际应用领域数据流中的噪声数据表现形式多样，如果事先对噪声进行形式化定义，对数据流中的噪音建模，有助于区分概念变化的类型与噪音。通常将噪声数据视为离群点(Outlier)，数据流的离群点检测也是数据流挖掘的重要内容，可作为数据流分类挖掘之前的预处理。

在数据流中，噪音不仅在概念变化之前会影响分类器的分类精度，而且会影响分类器对概念变化的适应。通常数据流分类器设计成自适应工作模式，如果分类算法对噪音过于敏感，就会错误地将噪音解释为概念变化；相反地，健壮的分类器会错误地将新概念的实例解释为噪音，因此收敛到新概念的速度较慢。文献[9]基于随机决策树，引入 Hoeffding Bounds 不等式来检测和区分概念变化和噪音，根据检测结果动态调整滑动窗口的大小和变化检测周期。

如何准确地检测到多种类型的概念变化(包括噪音区别、渐进式、突变式、数据分布变化以及多种复合类型等)，采取相应的对策使分类模型适应数据流的变化，这是数据流中概念变化研究中的一个重要问题；同时如何利用已有概念或已发现的变化特征去指导预测未来的变化趋势也是研究问题之一；另外，现实数据流中不可避免地有噪音干扰着概念发现与变化的检测，如何减少噪音的扰动，正确地从噪音中区分出真正的概念变化特征是另一个重要的问题。

数据流中概念发现的实质就是从含有噪声的数据流中提取被研究对象的数学模型，通常这个模型只是对象的特性在某种准则意义下(不确定性或概率)的一种近似，其近似的程度取决于人们对先验知识的认识深化程度和对数据集合性质的了解，以及所选用的概念发现方法与变化检测方法的合理性。现实世界的数据流中概念变化具有隐含、多变等特点，探索概念变化特征的形式化表示方法，将变化特征作为一种更粗粒度、更高层次的知识，用于概念发现及变化预测、检测过程，有助于实现定向数据流挖掘，提高数据流的知识发现效率。

3 概念变化的处理方法

数据流挖掘研究领域对概念变化已经进行了广泛地研究，通常的做法就是对概念变化做一定的假设，通过基于假设检测的方法来检测概念变化。要有效处理数据流中的概念变化，最为理想的办法是能够辨识概念变化将在哪个时间点发生，这样就能及时地对分类模型进行重新训练并更新。到目前为止，还没有实现可以适应不同概念变化的方案。

2004 年专刊[10]主要探讨了增量学习方法使已有分类器适应概念变化；随后出现大量，如自适应概念变化的分类器[11-13]，机器学习中的半监督学习和主动学习[14-17]等新方法应用到概念变化研究中。

3.1 自适应数据流概念变化的分类器

数据流分类的主要任务就是在存储空间比较有限和时间效率要求较高的情况下，设计出有效的算法来处理数据流的分类模型，并能实时响应用户的分类要求，尽可能地增强分类

的准确性，并能自适应(Aadptive)数据流分类过程中的概念变化问题。

数据流概念变化可以分为全局和局部概念变化，全局模型适宜处理稳定不变的数据流，而许多全局的急切式分类算法并不适宜处理局部变化的概念变化。而在现实数据流中，概念变化又很有可能发生在局部。如特殊类型的垃圾邮件可能随时间的变化而发生改变。当发生局部概念变化时，懒惰式分类器[11]由于其自身的局部特征却能很好地适应局部的概念变化。

根据数据流分类系统中所使用的分类器的个数，分类器可分为单分类器算法与集成分类器算法[1]。

集成式分类器是多个基础分类器通过某种评价机制对数据流中的样本进行综合评价的一种集成方法。集成分类器算法已经被证实在处理存在概念变化的数据流数据时，比单独的分类器具有更好的适应性和精确性，现有的方法包括贝叶斯平均、包装(Bagging)、推进(Boosting)。文献[12]提出了一种称为 M_ID4 的数据流挖掘算法，即通过尽量少的训练样本来实现概念变化检测的快速方法，其利用多分类器综合技术，实现了数据流中概念变化的增量式检测和挖掘，但是集成分类器算法在某些情况下还不能完全捕捉数据流中概念变化的变化。

3.2 基于样本的方法

(1)样本选择(Instance Selection)

样本选择指其目标是选择与当前概念相关的样本。基于样本选择的机制是天然的处理概念变化的方法。对于数据流，一般通过滑动窗口机制对样本进行选择，通常利用滑动窗口从最近到达的样本中选择样本子集，然后使用学习到的概念对紧接着到来的样本做预测。对数据流按块(Block)处理也是样本选择机制。

(2)样本权重(Instance Weighting)

样本权重的方法是基于以下的思想：一个样本的重要性应该随着时间而逐渐减弱。如在信息过滤系统中，用户对于某个主题的关注度可能随时间缓慢降低，这就很难找到一个准确时间点去判断用户对这个主题毫无兴趣。一旦确定训练集中所有样本的权重，就可采用某些能处理样本权重的学习算法进行下一步的学习。

(3)样本处理(Instance Processing)

根据数据到达方式的不同，可分为单样本到达概念变化处理方法与块状到达概念变化处理方法。采用何种方法完全取决于数据流入的方式与速度，以及应用问题的时间性要求，当然这两种方法是可以相互转换的。一般情况下，块状到达概念变化处理方法更适合于超大规模数据流的处理，而单样本到达概念变化处理方法对样本逐一进行处理，这可能相当消耗时间资源的工作。

3.3 数据流概念变化度量的标准

数据流上的概念变化处理需要解决两个基本问题：检测变化和修正模型，即当数据流中的数据不再遵从已有概念模型时，如何检测这种概念的变化；在检测到变化之后，如何更新当前的概念模型使其快速地收敛于新的概念。解决这两个基础问题的前提是必须度量概念变化的程度，也就是概念变化度量的标准问题。国内外研究者们在此领域已取得了一定的成果，但仍存在一些问题亟待解决。

文献[18]提出检测概念变化的两个标准：1)基于模型对新流数据正确预测的平均置信度；2)置信度低于给定阈值的事件百分比。真实数据流应用中的变化采用以上的标准面临的困难是变化比较缓慢，或者未发生根本性质的概念变化。因此度量两个概率分布的差异成为检测概念变化的关键，散度是很好的度量方法，常用散度有[6]：Kullback-Leibler divergence(K-L 散度)和 Jensen-Shannon divergence(J-S 散度)。K-L 散度适合于事件流(Event Streams)的变化检测，设事件流 S_m 中所有样本的采样来自同一分布 $P_m(x,y)$，下个事件流 S_n 的分布为 $P_n(x,y)$，Ω 是 x,y 的值集，则 $KL(P_m(x,y)\|P_n(x,y))=\int_{\Omega}P_m(x,y)\ln(P_m(x,y)/P_n(x,y))\geqslant 0$，其中 $P(x,y)=p(y|x)p(x)$。当 $P(x)$ 改变而 $P(x|y)$ 不变，称为特征改变(feature change)，也称虚拟概念漂移。通过分析概念漂移发生的原因，文献[6]把数据流上的概念漂移分为两种类型：松弛的概念漂移(仅仅样本的先验分布 $p(x)$ 发生变化)和严格的概念漂移(样本的先验分布 $p(x)$ 和潜在模式 $p(y|x)$ 同时发生变化)。

文献[19]提出了基于熵(Entropy)的概念变化检测算法，将一种熵的计算作为训练集之间样本分布的区别，由于熵非常适合度量系统的不确定性，因此也经常作为数据流分类器的决策依据[12]。

文献[20]依据统计学理论提出基于鞅(Martingale)的概念漂移检测方法，先综合考虑数据分布质心和半径改变引起概念的漂移，提出相异度量方法，然后对数据流采用双向统计的方法以更准确地标识数据分布并映射到均匀分布序列，最后计算双重随机幂鞅的均值，并利用停时定理来判断数据流中是否有概念漂移发生。

3.4 基于机器学习的概念变化检测机制

从机器学习的观点来看，数据流分类就是一个概念学习过程，它通过搜索训练数据集中蕴含的概念(分类规则)，进而预测未来到达数据的概念(分类规则)；目前大多数是采用统计方法来检测数据流中的概念变化，或者根据最新的数据来动态更新分类器以适应概念变化。如果当前的数据块和即将到来的数据块的概率分布相同或相似，这些方法无疑都是有效的。但在现实数据流中，这种假设并不一定成立。

机器学习的一个难题是现实世界的概念是不断变化的，它们通常依赖于一些潜在的上下文(hidden context)。一个典型的例子就是天气预报规则会因为季节不同而产生根本的变化[21]。通常变化原因是潜在的，不可能预先知道。潜在的上下文变化会或多或少地导致目标概念的变化，通常也被称作概念变化。

机器学习中主动学习与半监督学习方法也常常用于数据流分类中，在不存在概念漂移的情况下，半监督学习是数据流分类的最好方法[14]。主动学习也经常应用于数据流分类算法中[15]，在存在概念漂移的情况下，结合主动学习与半监督学习是比较好的数据流分类机制[17]。

3.5 预测周期性概念变化

从分类的角度来看，已有的数据流分类算法主要将发现新概念、提高分类算法的适应性作为研究重点，而如何组织和利用周期性出现的概念方面的研究还比较欠缺。由于所属领域的差异性，不同数据流中概念重现的周期存在一定的差异，若能将领域知识合理纳入组织周期性概念的考虑范围，就会提高算法处理周期性概念的能力。周期性更新的模型代价比较高，不一定适合真实的实际用途。解决的方法是先检测数

据流概念变化，仅在必要的时候进行分类模型的调整。

除了能够快速地识别和适应新概念外，一个完善的数据流分类系统还须具有一定的鲁棒性和识别周期性概念变化的能力。对一个变化的概念加上某些限制条件，如对数据流的背景做某些限制，理论上这样变化的概念是可以预测的。针对一些特殊类型的数据流，也有相关工作研究了如何适应其概念变化。

按照概念变化产生的形式，可以将其分为非周期性变化和周期性变化[22]，周期性概念变化就是相同或相似的概念周期性地出现，如天气的季节性变化。周期性数据流概念重现即待学习的目标概念(Target Concept)往往依赖于隐藏背景(Hidden Context)，日常经验表明许多隐藏的背景可能会重现。

定义 3(周期性概念变化) 数据流中数据在过去某个时间段形成的概念为 C'，经过概念变化后目前的概念状态为 C，经过一段时间(周期) T 后，C' 再次出现，并且周期性地再次转变为 C，这种现象称之为周期性概念变化。

针对部分特殊数据流存在周期性概念变化的特点，可以利用该数据流以前的概念(历史概念)所对应的数据流模型来预测再度出现的数据流，分类器可以更快地更新类型，加快数据流在线分类的速度[23]。

目前的数据流分类器一般在概念变化发生时，需要重新学习分类器，当应用于周期性概念变化的分析时，这些分类器没有充分利用"周期性"的特点，当以前的概念重新出现时，再次学习分类器代价是比较大的，也是没有必要的。

数据流分类主要关注于预测每个具体实例的类标记，而对于预测即将到来的概念的研究还未进行。只有概念变化发生后，概念变化才能被检测出来，相当于被动的概念检测。这种方法通常有一段较长时间的延迟，有可能延迟的时间内数据流又发生了概念突变，所以概念变化的预测准确度非常低，而提前预测概念变化有都着重要的意义。可以通过对历史概念的归纳与学习来预测未来的概念，其分如下 3 步进行：

1. 组织原始历史数据出现过的概念成为压缩历史概念，通过历史概念识别新概念及重复出现概念；

2. 从概念历史中学习概念的迁移模式；

3. 在两个层次上实现有效的预测，即对每个即将来临概念的一般层次预测和对每个实例类别的特殊预测。

概念之间的迁移可能是概率性而非确定性的，概念变换的过程可看作是马尔可夫链，一个概念对应马尔可夫链的一种状态，从概念变换的过程中学习概念转移的规律，从而得到表示概念变换规律的转移矩阵，并用历史概念来检验新的概念是否是历史概念的重现。

图 2 是周期性数据流的概念变化处理模型示意图，C1，C2，C3，C4 分别属于 4 种不同的概念模式，每当有新的概念出现时，历史概念集添加新的概念类型，当周期性数据流出现概念重现时，从历史概念集中直接调用相应的分类模式，从而大大加快了数据流分类的效率。

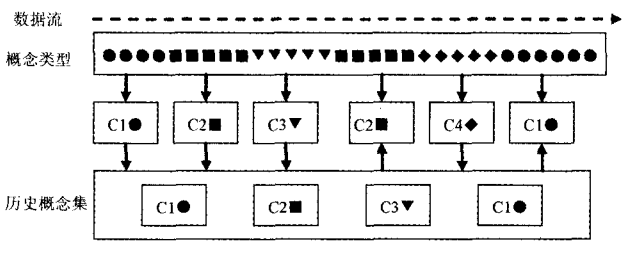

图 2 周期数据流中的历史概念

文献[24]提出的 RePro 算法把概念视为决策树所产生的分类规则或 Naïve Bayes 分类器所得到的概率表，将初始数据集生成历史概念(historical concept)，将概念变换的过程看作是马尔可夫链(Markov Chain)，一个概念对应马尔可夫链的一种状态，并从概念变换的过程中学习概念转移的规律，最终得到表示概念变换规律的转移矩阵，以此可以预测新的概念(new concept)是否是历史概念的重现。

上述是针对特定数据流的概念变化类型，但大多数数据流中发生的概念变化具有隐含、未知、易于变化以及多重性。不同变化的变化模式会交替出现。原有的模型是在发生概念变化之后采取的反应措施，对未来的概念变化不能及时进行预测。利用已有概念或已发现的概念变化特征去预测将来的变化趋势能更好地适应概念变化的多重性。

4 数据流概念变化研究新问题及相应对策

目前，虽然已经提出了一些解决数据流中概念变化的方法和机制。然而，总体说来，数据流中的概念变化处理还处于理论和方法的探讨阶段，还没有一种方法可以完全地检测到数据流中的所有概念的变化。同时随着数据流挖掘的研究深入发展，新的数据流概念变化研究的难点也不断地被发现：

1. 多种概念同时并存，且类别严重不均衡

多种概念同时并存的现象发生，使得训练数据集中蕴含的映射模型不再是唯一的，而是混杂着多种映射模型。而常见的数据流分类算法常常是以假设唯一的映射模型为前提的，如何在概念变化情况下保证分类精度就比较困难。

对应策略：采用多标签数据流分类方法。多标签数据流分类研究目前还处于刚刚起步阶段，文献[25,26]针对概念漂移(Concept Drift)和类不均衡(Class Imbalance)的数据流，采取双重滑动窗口机制，提出基于 KNN 算法的数据流多标记分类框架，并针对特定类型的数据流进行相关优化算法的设计。

2. 如何突出最新最近概念的重要性

一般来说当前流入的数据最有价值，数据流分类挖掘算法需要做到能够尽可能及时地发现蕴含在当前数据流中的最新映射模型[27,28]。使用旧的映射模型来对新到达的数据进行类别判定，往往有失准确。

对应策略：基于遗忘机制的数据流处理机制。

先前的流值对未来预测值的影响不同，一般越接近预测点的流值对预测流值的影响越大，相隔的时间越长，可能存在的误差越大，影响程度也越小。如果概念变化发生速度较快并且以不一致方式发生时，概念更新时间被延迟，则可能造成相关信息丢失。难点在遗忘机制中的遗忘系数如何定义，可以利用机器学习的方法自适应地动态调整遗忘系数，从而与概念变化的处理能保持一致。

3. 如何自适应不同类型的概念变化

事实上很少有固定概念变化类型的数据流，通常数据流都有多种概念变化的类型，如何能自适应不同类型的概念变化，是数据流分类器的最大的挑战，严格来讲，不存在一种能自适应不同概念变化类型的数据流分类器。

对应策略：集成分类器是最佳选择，但集成式-分类器速度较慢，如何提升集成式分类的工作效率是将来的研究方向。此外，结合离线分类器与在线聚类也是不错的选择[29]。

(下转第 386 页)

[11] Reis G A, Chang J, Vachharajani N, et al. SWIFT: Software implemented fault tolerance[OL]. http://liberty.princeton.edu/publications.cg03-swift.pdf

[12] Nicolescu B, Savaria Y, Velazco R. Software Detection Mechanisms Providing Full Coverage Against Single Bit-Flip Faults [J]. IEEE Transactions on Nuclear Science, 2004, 51(6): 3510-3518

[13] Li Ai-guo, Hong Bing-rong. Software implemented transient fault detection in space computer[J]. Aerospace Science and Technology, 2007, 11(2/3): 245-252

(上接第 350 页)

结束语 数据流的时变性决定了分析数据流中概念变化的重要性。目前概念变化的研究非常活跃，一些解决数据流中概念变化的方法和机制已经被提出。然而，总体说来，数据流中的概念变化处理还处于理论和方法的探讨阶段，还没有一种方法可以完全地检测到数据流中的所有概念的变化。为了简化概念变化的处理，通常采取的做法有：假设数据中只存在某种类型的概念变化；数据流的数据分布恒定；同时限定概念变化的速率和范围等。而真实的数据流中往往不符合以上 3 个条件，还需要进一步研究其概念变化本质。

目前数据流研究逐渐面向不确定数据流(Uncertain Data streams)[30]，如移动位置流(Location Streams)、文本数据流(Text Streams)、传感数据流(Sensor streams)，这些数据流具有不确定性(Uncertainty)，概念变化研究仍有着重要的应用价值，可以将概念变化与异常检测相关结合，以及时有效地发现数据流的异常现象。不确定数据流中的概念变化研究仍然有着重要的意义。

参 考 文 献

[1] 王涛，李舟军，颜跃进，等. 数据流挖掘分类技术综述[J]. 计算机研究与发展，2007，44(11)：1809-1815

[2] Tsymbal A. The problem of concept drift: definitions and related work. TCD-CS-2004-15 [R]. Dublin, Ireland: Department of Computer Science Trinity College, Trinity College, 2004

[3] Gama J. A survey on learning from data streams: current and future trends[J]. Progress in Artificial Intelligence, 2012, 1(1): 45-55

[4] 曾华军. 机器学习[M]. 张银奎，等译. 北京：机械工业出版社，2003

[5] 王永利. 基于概要的数据流挖掘若干研究[D]. 南京：东南大学，2006

[6] 张鹏. 挖掘概念漂移的数据流[D]. 北京：中国科学院，2009

[7] Widmer G, Kubat M. Effective learning in dynamic environments by explicit context tracking[C]//Proceedings of the Sixth European Conference on Machine Learning. Vienna, Austria, 1993: 69-101

[8] Schlimmer J, Granger R. Incremental learning from noisy data [J]. Machine Learning, 1986, 1(3): 317-354

[9] Li P-P, Wu X-D, Gao Y-J. A Random Decision Tree Ensemble for Mining Concept Drifts from Noisy Data Streams[J]. Applied Artificial Intelligence, 2010, 24(7): 680-710

[10] Kubat G, Gama J, Utgoff P. Special issue on incremental learning systems capable of dealing with concept drift[Z]. Amsterdam, Netherlands: Intelligent Data Analysis, 2004

[11] 尹志武，黄上腾. 一种自适应局部概念漂移的数据流分类算法[J]. 计算机科学，2008，35(2)：138-139

[12] 孙岳，毛国君，刘旭，等. 基于多分类器的数据流中的概念漂移挖掘[J]. 自动化学报，2008，34(1)：93-97

[13] 刘耀宗，张宏，王永利. 一种自适应概念变化的数据流分类器[J]. 计算机研究与发展，2007，44(Suppl)：63-68

[14] Klinkenberg R. Using labeled and unlabeled data to learning drifting concepts[C]//Proceedings of the Workshop notes of the 1 JCAI-01 Workshop on Learning from Temporal and Spatial Data. Menlo Park, CA, USA, 2001: 16-24

[15] Zhang P, Zhu X, Guo L. Mining data streams with labeled and unlabeled training examples[C]//Proceedings of the ninth IEEE International Conference on Data Mining. Miami, FL, USA, 2009: 627-636

[16] Widmer G, Kubat M. Effective learning in dynamic environments by explicit context tracking[C]//Proceedings of the Sixth European Conference on Machine Learning. Vienna, Austria, 1993: 69-101

[17] Kholghi M, Keyvanpour M R. Active learning framework combining semi-supervised approach for data stream mining[J]. Intelligent Computing and Information Science, 2011, 135: 238-243

[18] Hulten G, Spencer L, Domingos P. Mining time-changing data streams[C]//Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, 2001: 97-106

[19] Peter V, Abranham B. Entropy-based concept drift detection[C]// Proceedings of the 6th International Conference on Data Mining. Hong Kong, China, 2006: 1113-1118

[20] 张育培，柴玉梅，王黎明. 基于鞅的数据流概念漂移检测方法[J]. 小型微型计算机系统，2013，34(8)：1787-1792

[21] Verpeck J T, Meehl G A, BonyS, et al. Climate data challenges in the 21st century[J]. Science, 2011, 331(6018): 700-702

[22] 罗秀，王大玲，冯时，等. 一种面向周期性概念漂移的数据流分类算法[J]. 计算机研究与发展，2009，46(Suppl)：400-405

[23] Li P P, Wu X D, Hu X G. Mining recurring concept drifts with limited labeled streaming data[C]//Proceedings of the 2th Asian Conference on Machine Learning. Tokyo, Japan, 2010: 241-252

[24] Yang Ying, Wu Xin-dong, Zhu Xing-quan. Combining proactive and reactive predictions for data streams[C]// Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, Illinois, USA, 2005: 710-715

[25] Chen S, He H, Li K, et al. MuSeRA: multiple selectively recursive approach towards nonstationary imbalanced stream data mining[C]//Proceedings of the 2010 International Joint Conference on Neural Networks. Barcelona, Spain, 2010: 1-8

[26] Chen S, He H. SERA: selectively recursive approach towards nonstationary imbalanced stream data mining[C]//Proceedings of the 2009 International Joint Conference on Neural Networks. Atlanta, Georgian, USA, 2009: 522-529

[27] 辛铁，郭躬德，陈黎飞，等. IKnnM-DHecoc：一种决概念漂移问题的算法[J]. 计算机研究与发展，2011，48(4)：592-601

[28] 朱群，张玉红，胡学钢，等. 一种基于双层窗口的概念漂移数据流分类算法[J]. 自动化学报，2011，37(9)：1078-1083

[29] 杨春宇. 数据流上聚类与分类算法[D]. 北京：清华大学，2009

[30] 周傲英，金澈清，王国仁，等. 不确定性数据管理技术综述[J]. 计算机学报，2009，32(1)：1-16