演化数据的学习

张长水 张见闻

(清华大学自动化系 智能技术与系统国家重点实验室 信息科学与技术国家实验室(筹) 北京 100084)

摘 要 在一些实际问题中,数据的分布随时间的变化而逐渐变化,这类数据的学习问题被称之为演化数据的学习.文中综述了演化数据上的学习方面的研究进展.提出了今后需要关注的一些问题,如数据演化的机制、一般性的假设问题、演化数据分类等等.

关键词 机器学习;演化数据;非监督学习;半监督学习

中图法分类号 TP18 **DOI号** 10.3724/SP.J.1016.2013.00310

Learning on Time-Evolving Data

ZHANG Chang-Shui ZHANG Jian-Wen

(*State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing* 100084)

Abstract In some real applications, the data distributions often evolve over time. Machine learning on this kind of data is named learning on time-evolving data. This paper surveys the recent advance of the research on this topic, and introduces some problems worth attention in the future, such as the mechanism of Time-Evolving Data, generality, classification.

Keywords machine learning; time-evolving data; unsupervised learning; semi-supervised learning

1 引 言

机器学习问题可以直观地表述为,从经验样本集中训练学习机器,使之能尽可能好地拟合或预测未见的测试样本集.传统的统计机器学习的一个基本假设是训练数据和测试数据都是从一个概率分布中独立地抽取得到的[1].而一些实际应用问题的数据分布是随时间动态变化的,例如,在新闻、博客和BBS等在线媒体中,人们讨论的话题大多数都会随时间发生变化,即使对于同一个话题,一年前和当前的内容也不完全相同,比如"时尚"、"高新技术"等.这在互联网数据分析中被称为概念漂移.

这样一类数据分布动态变化的问题给机器学习带来一些深刻的变化和挑战.一方面,就拟合或预测未来数据而言,由于独立同分布假设显然不成立,我们不能像对待传统的学习问题那样,把在历史数据上训练得到的学习机器直接作用于未来的数据,传统的很多理论和方法都需要修正.另一方面,从建模的角度,缺少独立性和同分布性,样本集的概率不能简单地再写成各样本概率的乘积.最后,日益丰富的应用问题中,人们不仅需要学习机器能很好地拟合或预测未来数据,同时也希望它能够揭示出数据的动态演化规律,从而让人们可以更好地理解数据.例如,在网络舆论分析中,用户不仅关心每一时刻的主要讨论内容,同时也关心这些内容的动态模式,如某个话题的产生、传播、变化和消亡等.这就区别于传统的学习问题,为机器学习带来一个新的任务——学习数据的动态演化机制.传统的学习方法归根结底是对某一静态的数据分布的学习,没有提供学习数据分布的变化规律的办法.

这一问题逐渐引起机器学习和数据挖掘领域的

收稿日期:2011-09-05;最终修改稿收到日期:2012-11-18.本课题得到国家自然科学基金(61075004)、浙江省自然科学基金(Y1110661)资助.**张长水**,男,1965年生,博士,教授,博士生导师,研究领域为机器学习、模式识别、计算机视觉、数据挖掘、复杂网络.**张见闻**,男,博士,研究方向为机器学习、数据挖掘及应用.

重视[2-12]，并将分布随时间变化的数据称为非平稳数据(nonstationary data)或演化数据(time-evolving data，evolutionary data)①. 演化数据上的学习，已经成为机器学习和数据挖掘中一个新的重要研究问题.

在演化数据的学习问题中，聚类问题的研究有其特殊的意义，与监督问题相比，吸引了更多的研究兴趣. 传统的聚类算法，如 K 均值[13]和谱聚类[14-15]，处理的是静态的数据，算法被要求在给定的数据集合上有尽可能好的聚类划分. 演化数据上的聚类是这样一个学习任务：数据的分布随时间而变化，在每一时刻，新的数据进入系统，系统要求为这一批数据作出聚类划分. Chakrabarti 等人[6]2006 年第 1 次明确地提出这一问题，并称之为演化聚类. 已有的工作指出，演化聚类有 3 个方面的目的[6,16-17]：首先，每一时刻上的数据聚类性能要尽可能好；其次，我们希望通过聚类发掘数据的演化机制，例如聚类的出现、变化、分裂、消失等；最后，聚类结果在时间上要尽可能平滑.

直观上，我们可以尝试用传统的聚类方法以两种方式来解决演化聚类问题. 第 1 种，忽略数据性质随时间的变化，在随时间累积的总体数据上直接应用传统的聚类算法. 但是，当数据分布随时间发生变化时，即使在每一时刻上的聚类是明显的，整体数据上的聚类可能也是毫无意义的. 这样在随时间积累的总体数据上直接应用传统的聚类算法是没有意义的. 第 2 种，忽略不同时刻数据之间的关联性，在每一时刻的数据上独立地应用传统的聚类算法. 这会导致两方面的问题. 一方面，保持聚类结果随时间的平滑性这一目的不能达到. 例如，对于依赖初始化的一些算法，如 K 均值聚类和高斯混合模型等，由于局部极值的存在，即使相邻时刻的数据分布很接近，相邻时刻上的聚类结果也可能会相距甚远. 另一方面，在实际应用中，相比于样本维数，样本量往往不够，完全抛弃邻近时刻的数据很可惜. 值得说明的是，在数据演化机制的学习上，用上述两种传统的途径也不能达到目的.

下面以一个例子来说明上述问题. 我们构造了一个随时间演化的高斯混合模型，它包含有 3 个高斯成分，其参数随时间而缓慢变化(共 30 个时刻)，在每一时刻上，样本从该时刻的混合模型中抽取出来，抽取的过程带有噪声，如图 1 所示. 图 1 中，每一个高斯成分的样本用一种符号(三角，方块，圆圈)表示. 该数据分布的变化十分明显，以至于将所有时刻的数据累积起来时，不同的类别已经完全混叠在一起了(图 1(e)).

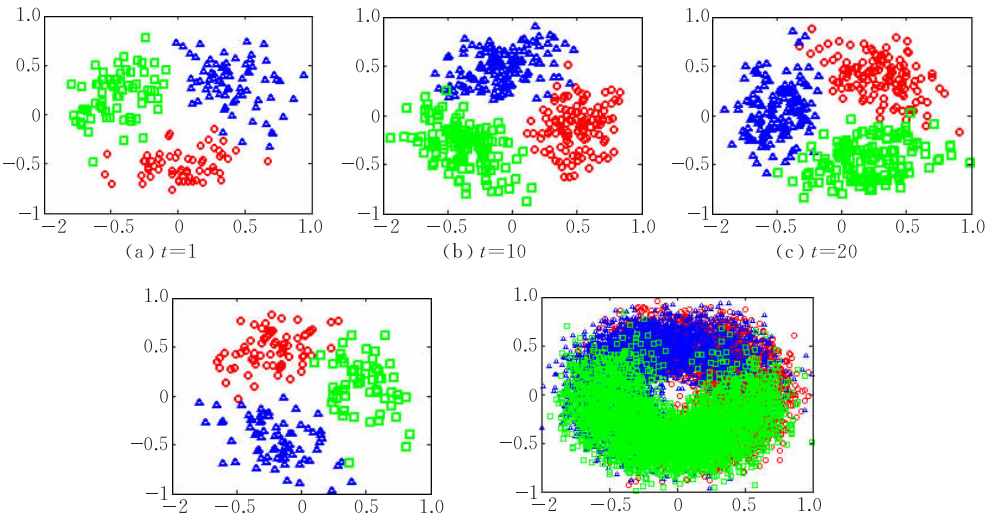


图 1　参数随时间缓慢变化的带噪声高斯混合模型产生的数据

演化数据的学习存在如下几个主要的难点：

(1) 演化数据背后的统计假设还不明确，对“演化”行为的本质还缺乏清晰的描述，这使得我们缺乏方法研究和建模的基础.

(2) 如果引入某种新的假设，在传统的独立同分布假设下的最大似然、最大后验、贝叶斯估计等方法在演化数据中如何应用.

(3) 复杂的演化模式的学习. 例如，演化数据上通常存在类别数的变化、类别的出现、变化和消失等；另外，实际应用可能关心在多个相关联的数据源之间是否存在关联演化模式，如不同的 BBS 站点之

① “非平稳数据”一词揭示数据的统计性质，“演化数据”则是更为直观地描述数据的变化. 我们遵从大多数应用研究的称谓，称之为“演化数据”.

间、新闻与 BBS、博客之间的关系等.对这些复杂的演化行为建模和学习,都是很困难的具体问题.

(4)传统的聚类模型和算法多种多样,能否找到一种途径,简单有效地将传统的一些算法推广应用到演化数据上.

2 几个相关的研究问题

在具体介绍演化聚类的研究工作之前,我们首先介绍几个与演化数据学习紧密联系的研究问题.

在机器学习和数据挖掘领域,有几个比较重要的研究内容:增量式学习(incremental learning)、数据流上的学习(learning on data stream)和时间序列上的学习,它们和演化数据上的学习在某些方面有相似之处或紧密联系,但存在本质区别.

增量式学习是为了将传统的学习算法应用在超大规模数据上的一个途径.在增量式学习中,数据被要求组织成批的形式或数据流的形式,这是出于对一些实际应用因素的考虑.例如,数据太大不能一次性载入内存,或者数据本身是在线搜集的,算法也需要在线地运行.增量式学习希望通过对数据的一次性扫描方式(single-pass)[18],能够达到在整体数据上离线学习的同样效果.和演化学习不同的是在增量式学习中,(1)学习结果等价于整批数据上的学习;(2)没有强调数据的时间性质;而相反,增量式学习追求的是算法对数据到达的顺序不敏感.因此,增量式学习并不针对统计性质随时间变化的数据,而只是对静态数据的一种操作方式,以获得和原始方式下相同(或相近)的性能.

数据流上的学习[4,19-22]针对是连续高速产生的数据流.因此,将所有数据都存储下来是不现实的,同时,"流"的特性不允许算法有多次扫描全体数据的机会.早期的关于数据流上的聚类的研究[21]把它作为一种单次扫描(single-pass)聚类,就和增量式聚类相同.而最近的研究则开始强调数据的时间变化特性,算法只限制在某个时间区间内[22].与演化数据学习不同的是:首先,数据流上的学习并不强调学习结果在时间上的平滑性,而在演化数据学习中,这是一个主要的目标;其次,在数据流学习中,每一时刻只有一个样本到达,算法要求的是在某个数据窗内能够快速的计算.而在演化数据学习中,某一时刻的数据是来自于与当前时刻对应的某一分布的一批样本,强调的是利用时间上下文信息帮助提高学习性能、保持学习结果的时间平滑性以及学习数据演化的规律.

时间序列上的学习(learning on time-series)的基本数据是一个时间序列,或者一个时间序列片段,而在演化数据学习中,会把时间序列某个时刻的数据作为一个样本.

3 演化数据的学习方法归类

目前针对演化数据学习的研究多集中在演化数据的聚类方面.这些研究可以从以下几种不同的角度来分类归纳.

3.1 在线式与离线式

从算法对数据的可见性,演化聚类算法可以分为在线式和离线式.在对第 t 时刻的数据做聚类时,在线式的算法只能利用 t 时刻及以前的数据,而离线式的算法可以利用包括 t 时刻以后的所有时刻的数据.虽然我们在介绍演化数据的概念时,是以实际的在线式应用引入的,但演化聚类的离线式算法是有其应用意义的,例如离线式的网络舆论分析、为数据可视化提供的离线式聚类分析[23]等.演化聚类算法的在线式算法更贴近实际应用的情形;离线式的算法则更利于对数据演化行为的挖掘.

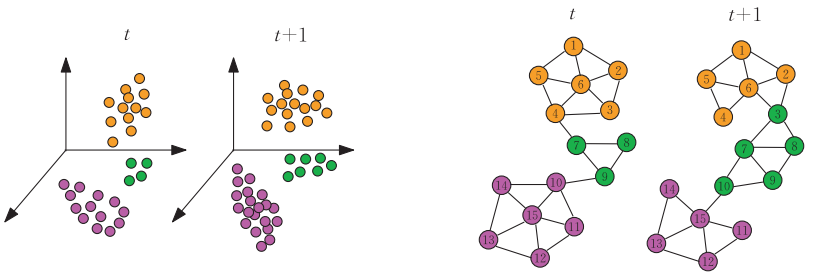
3.2 参数化与非参数化

聚类数目的确定是一个模型选择问题.在传统的聚类问题中,原则上可以采用一些经典的模型选择方法,如 Akaike 信息准则(AIC)[24]和贝叶斯信息准则(BIC)[25]等,不过这些方法在实际应用中难以使用,更多的情况还是依靠用户对数据的先验而人为指定.在演化聚类中,这一问题显得更重要,一方面,演化聚类中不同时刻的聚类数可能会变化,这给用户指定合适的聚类数带来更大的难度;另一方面,聚类数的变化还蕴含了不同时刻间的聚类类别对应关系的问题.我们即使使用某种准则为每一时刻挑选聚类数,聚类之间的对应关系还是需要另外的机制来确定.因此,在贝叶斯框架下,有一类探讨如何从数据中推断聚类数目的非参数方法[26]在演化聚类的研究中得到了广泛的应用.这一类方法在先验中不限定聚类数目,通过在不同的时刻之间跟踪模型的增长过程又可以方便地建立不同时刻聚类之间的对应关系.与此对应,根据指定的聚类数目来学习模型的方法,则称为参数化方法.

3.3 两种数据形式

在演化数据的研究过程中,出现了两种不同的

数据特征形式. 图 2 给出了一个简单的示例，其中每一个圆圈表示一个数据样本. 第 1 种和传统的学习问题相同，数据样本被表示为共同的有限维特征空间中的向量，如图 2(a)所示. 每一时刻的样本是从该时刻对应的数据分布中抽取出来的，不同时刻之间的样本没有对应关系. 第 2 种是关系型数据，数据样本没有自身的特征表示，而只有样本之间的链接关系，如图 2(b)所示，图中圆圈的编号唯一地标识一个样本，圆圈之间的连线表示样本之间的链接关系(这里没有考虑权重). 这样的数据实际上构成一个图，图的结点就是一个样本点. 而随时间推进，结点之间的链接关系会发生变化，之前存在的链接可能消失，之前没有的链接可能建立. 这两种数据形式分别对应不同的具体应用问题. 例如，对于网络舆论分析这样的应用，不同时刻的文本之间并没有确定的对应关系，随时间变化的是数据的统计性质，反映的是网络舆论的兴趣或热点的变化. 而对于像社会网络分析这样的应用而言，如果不考虑新的结点的加入，在不同的时刻结点群是不变的，变化的只是结点之间的链接关系.



(a) 特征空间中的概率分布随时间变化　　(b) 结点之间的链接关系随时间变化

图 2　演化聚类中的两种数据形式

3.4　对演化机制的两种建模途径

从对数据演化行为的建模手段来看，一些工作假设了数据演化的产生式随机过程，对数据演化机制进行显式地建模；还有一些工作不对演化过程作具体的产生式假设，而是通过“黑盒”式地引入时间正则项来表达数据统计性质的变化给模型带来的影响.

下面我们将按照对演化机制建模的不同途径来分别介绍具体的相关方法.

4　研究工作现状

4.1　基于平滑性正则的方法

传统的聚类问题表达为最小化一个目标函数时，如果该目标函数并不依赖于数据的生成模型，而且不考虑聚类模型对于未来样本的预测，就算法而言，这一类聚类方法对独立同分布假设没有直接的依赖关系. 在演化数据上，一个最直接的处理方法就是，在每一时刻的数据上，采用传统聚类方法的原则设计损失函数，然后通过加入某种与时间有关的正则项来体现数据的动态演化对模型的影响. 这种方式的着眼点在于提高每一时刻的聚类质量和聚类结果的平滑性上，避开了对数据演化过程的复杂假设. Chakrabarti 等人[6]首先提出了这样一种非常一般性的在线式框架，将演化聚类每一时刻 t 上的任务建模为最大化下面的目标函数：

$$sq(C_t, M_t) - cp \cdot hc(C_{t-1}, C_t) \qquad (1)$$

其中，C_t 表示的是当前时刻的聚类模型；M_t 表示算法所利用到的当前时刻数据的信息；$sq(C_t, M_t)$ 衡量了算法在当前时刻数据上的聚类质量，而 $hc(C_{t-1}, C_t)$ 则表示算法的时间损失；参数 cp 是这两者之间的权衡系数. 这种框架只是一种一般的原则，并不能导出一般性的通用算法. 在该框架的指导下，针对 K 均值和分级聚类这两种比较常用的静态聚类算法，他们分别提出了具体的演化 K 均值算法和演化分级聚类算法. 以演化 K 均值为例，在该算法中，每一个聚类都被匹配到上一时刻距离最近的那个聚类，把所有这种配对的聚类之间的距离相加作为时间损失. Chi 等人[16]指出，这种贪婪的最近匹配方法可能不稳定，会对聚类中心的小的扰动十分敏感. 他们也利用类似文献[6]的思想，对谱聚类进行扩展，提出两种演化的谱聚类算法[16]，而谱聚类的扩展则不存在上述不稳定的问题. Tang 等人沿着演化谱聚类[16]的思想进一步工作，推广到多关系数据(multi-relational)聚类的情况，以处理动态的多关系社会网络中的社区(community)划分问题[27]. Wang 等人根据非负矩阵分解[28-30]和聚类之间的关系，采用了低秩矩阵逼近的方法来处理演化聚类的问题[31]. 他们将 t 时刻样本的聚类指示矩阵作用于 $t-1$ 时刻逼近 $t-1$ 时刻数据矩阵的误差作为时间损失项. 文

献[6]和传统的聚类方法一样,是基于数据样本的特征表示;而文献[16,27,31]处理的则是另外一种形式,即关系型数据.此外,在上述文献中,文献[6,31]给出的是在线式的方法,文献[16,27]同时给出了算法的在线和离线两种形式.文献[32]提出了通过时间平滑性正则利用历史信息来估计演化指数族混合模型的在线式途径.该途径将历史信息的来源归纳为两种:历史数据和历史模型,相应地导出了两种在线式算法.该途径指明了演化聚类和混合模型密度估计之间的关系,阐明了演化聚类的统计假设.真实文本数据上的实验表明,所提出的两种算法能同时提高每一时刻聚类性能和模型随时间的平滑性.

在这一类基于平滑性正则的方法中,平滑性正则项表达的含义可以归纳为两类:第 1 类是当前时刻的模型作用于前一时刻的数据带来的损失;第 2 类是直接表达为相邻两个时刻的模型之间的差异.在上述方法中,文献[16]分别设计了这两种类型的正则项;文献[31]属于第 1 类,文献[6,27]属于第 2 类,文献[32]综合考虑了这两类.到目前为止,这一类方法虽然可以允许不同时刻的聚类数不相同,但都需要用户预先指定每一时刻的聚类数,属于参数化方法.

文献[33]研究的是演化数据上的半监督学习问题:在已给历史数据的基础上,如何利用当前数据中少量的标注样本提高数据的分类性能.该文假设,每一时刻,有一批样本到来,其中有一部分已被标注.作者给出了基于可再生核希尔伯特空间(RKHS)的一个一般性的半监督学习框架.在该框架中,要求学习算法在每一时刻给出一个分类函数,使得在所有时刻上累积的分类损失最小.在互联网论坛的数据实验表明,其算法可以比传统的算法的性能有大幅度的提高.而这一工作也隐含地假设了数据的光滑性质.由于数据在演化,数据的独立同分布假设不再成立,因此,演化数据上的监督学习就没有存在的基础了.

4.2 对数据演化机制显式建模的方法

平滑性正则方法的优点是模型简单,建模和求解相对容易,但不足以表达数据的演化行为,例如,类别的出现、变化和消失等.因此,有一些对数据演化机制进行显式建模的方法被提出.对于演化数据,随时间变化的是数据分布,因此,演化机制的建模包含了两个方面的内容,一方面是每一时刻上数据的产生机制,另一方面是不同时刻模型的变化机制.这两方面的建模思路和传统问题中的方式都有差别.

对于第 1 个方面,在传统的非监督问题里,有限混合模型是最常用的建模手段,例如高斯混合模型.在演化数据上,虽然我们也可以沿用这一方式,在每一个时刻上采用一个高斯混合模型对该时刻的数据建模,例如 Wang 等人[34]的方法,但以此为基础我们难以找到处理聚类数量变化的方法.因此有一些研究工作探索基于无限混合模型来设计聚类数量变化的机制,例如文献[9,17,35-36].

文献[37]提出了一种从多个相关演化子集中挖掘复杂演化模式的贝叶斯非参数模型,并给出了基于吉布斯采样的贝叶斯推理算法.这一方法能够发现多个关联演化子集中的复杂演化模式,包括聚类的出现、变化、消失以及在不同子集之间的传播.而且,在该方法中,所有的聚类数都是从数据中自动学习,不需要人为指定.将该方法应用于对新闻、博客和讨论版三种典型的在线文本数据的分析,发现了一些有趣的关联演化模式.另外,文献[38-39]将演化聚类问题抽象为多任务聚类,提出了两种一般性的任务正则来表示任务之间的关系,相应地导出了两种基于 Bregman 散度的多任务聚类方法.这两种方法适用于多种数据散度的假设.大量真实文本数据集上的实验表明,两种方法能提高各任务聚类性能和各任务模型之间的一致性.他们进一步指出,这两种多任务聚类方法可以作为演化聚类的一般性框架,在具体参数设置下,它们是前面提出的两种演化聚类模型的等价或变体模型.

对于第 2 个方面,通常考虑模型的动态变化最常用的方式是马尔可夫跳转模型,但演化数据不同于时间序列,随时间动态变化的是数据的概率分布,表现为聚类模式的变化,因此应用马尔可夫跳转模型的难点在于如何定义“状态”以及不同时刻之间的转移矩阵.Wang 等人[34]和 Xu 等人[35]采用了马尔可夫跳转模型.在 Wang 等人的模型[34]中,每一时刻的数据用一个混合高斯模型来建模,并认为在所有的 T 时刻上,这些高斯模型只有有限的 K 种参数取值(通常认为 $T>K$),即在所有时刻上,最多只有有限的 K 种不同的高斯混合模型,而在时间上,这 K 种模型状态之间通过一个 $K\times K$ 的状态转移矩阵跳转,这一模型可以看作是一种隐式半马尔可夫模型(hidden semi-Markov models)[40-41].由于所有时刻的聚类模型只允许取有限的“状态”,这一模型适用于那些有明显的周期性且只有为数不多的“聚类状态”的情形.在对每一时刻的数据建模上,该方法依然使用了传统的有限混合模型,需要用户指定

每一时刻的聚类数目，属于参数化方法. 文献[35]则在不同时刻的聚类之间直接建立马尔可夫跳转矩阵，以表达从当前时刻到下一时刻聚类之间的对应关系. 该方法属于非参数方法. 这一方法的缺点是模型太复杂. 在实际演化数据上不同于传统的时间序列问题，聚类之间的相互跳转模型并不必要，增加了模型的复杂性，大多数情况下只是某个聚类一直往前变化或者消失. 因此，Ren 等人[9]、Ahmed 和 Xing[17]以及 Xu 等人[36]没有假设聚类之间的跳转，而是通过狄利克雷过程建立一个无限混合模型，其参数随时间而变化. 其中，Ren 等人[9]和 Xu 等人[36]只对混合系数的变化建模，而 Ahmed 和 Xing[17]同时还对混合成分本身的变化进行了建模. 由于他们也利用了狄利克雷过程，可以从数据中推理聚类数，都属于非参数方法. 上述这些对数据演化机制显式建模的方法处理的数据形式都是传统的特征表示的形式，也都是离线式算法.

此外，除聚类问题之外，针对文本的话题建模(topic modeling)这一具体问题，Blei 和 Lafferty[7]将静态情况下的话题模型[42]推广到动态情况，该问题不涉及话题数量的变化以及话题之间的关联演变行为. Wang 等人[43]则考虑了数据序贯式到达的连续时间的情况，利用布朗运动建模，并利用文本的稀疏性给出了一种高效的变分推理方法.

4.3　利用辅助信息的聚类

从机器学习的角度来抽象地考虑，演化聚类本质上可以看作是在各个时刻上利用相关的辅助信息来提高性能的聚类问题. 利用辅助信息的学习是机器学习中一个很宽泛的问题，它们解决问题的方法对演化聚类的研究具有借鉴意义. 这样的方法包括多任务聚类、迁移学习、聚类融合、多视图聚类等. 这方面的工作很多，由于篇幅原因，在此不再赘述.

5　总　结

演化数据上的学习方面的研究工作还不多，还有一些问题需要研究. 例如：数据演化的机制是什么？对演化数据是否存在一般性的假设从而让我们能更好的分析数据？考虑数据演化总是有用的和必要的吗？是否可以提供某种自适应机制以确定演化机制的影响？在(半)监督问题中，如果出现了类别数的变化怎么办？如何对演化数据更有效的分类. 对这些问题的关注和研究会使得演化数据上的学习前进一大步.

参　考　文　献

[1] Vapnik V N. The Nature of Statistical Learning Theory. New York, NY, USA: Springer-Verlag, 1995

[2] Bartlett P, Ben-David S, Kulkarni S. Learning changing concepts by exploiting the structure of change. Machine Learning, 2000, 41(2): 153-174

[3] Hulten G, Spencer L, Domingos P. Mining time-changing data streams//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). New York, 2001: 97-106

[4] Gaber M M, Zaslavsky A, Krishnaswamy S. Mining data streams: A review. ACM Sigmod Record, 2005, 34(2): 18-26

[5] Webb G, Ting K. On the application of roc analysis to predict classification performance under varying class distributions. Machine Learning, 2005, 58(1): 25-32

[6] Chakrabarti D, Kumar R, Tomkins A. Evolutionary clustering//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). New York, NY, 2006: 554-560

[7] Blei D M, Lafferty J D. Dynamic topic models//Proceedings of the 23rd International Conference on Machine Learning. New York, 2006: 113-120

[8] Bifet A, Gavalda R. Learning from time-changing data with adaptive windowing//Proceedings of the SIAM International Conference on Data Mining, Minneapolis. Minnesta, USA, 2007: 443-448

[9] Ren L, Dunson D B, Carin L. The dynamic hierarchical dirichlet process//Proceedings of the 25th International Conference on Machine Learning. New York, 2008: 824-831

[10] Bifet A, Holmes G, Pfahringer B. Leveraging bagging for evolving data streams//Proceedings of the 2010 European conference on Machine learning and Knowledge Discovery in Databases: Part I. Berlin, 2010: 135-150

[11] Ryabko D. On finding predictors for arbitrary families of processes. Journal of Machine Learning Research, 2010, 11: 581-602

[12] Ryabko D. Clustering processes//Füurnkranz J, Joachims T eds. Proceedings of the 27th International Conference on Machine Learning (ICML-10). Haifa, 2010. Israel: Omnipress, 2010: 919-926

[13] Hartigan J, Wong M. A k-means clustering algorithm. Applied Statistics, 1979, 28(1): 100-108

[14] Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems, 2002, 2: 849-856

[15] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[16] Chi Y, Song X, Zhou D et al. Evolutionary spectral clustering by incorporating temporal smoothness//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). New York, 2007: 153-162

[17] Ahmed A, Xing E. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: With applications to evolutionary clustering//Proceedings of the SIAM Conference on Data Mining. Atlanta, 2008: 219-230

[18] Gupta C, Grossman R. Genic: A single pass generalized incremental algorithm for clustering//Proceedings of the SIAM Conference on Data Mining. Nashville, TN, USA, 2004: 147-153

[19] Guha S, Mishra N, Motwani R et al. Clustering data streams//Proceedings of the 41st Annual Symposium on Foundations of Computer Science. Washington, DC, 2000: 359-366

[20] Barbará D. Requirements for clustering data streams. ACM SIGKDD Explorations Newsletter, 2002, 3(2): 23-27

[21] O'Callaghan L, Mishra N, Meyerson A et al. Streaming-data algorithms for high-quality clustering//Proceedings of the 18th International Conference on Data Engineering. San Jose, CA, USA, 2002: 685-694

[22] Aggarwal C C, Han J, Wang J et al. A framework for clustering evolving data streams//Proceedings of the 29th International Conference on Very Large Data Bases- Volume 29. Berlin, Germany, 2003: 81-92

[23] Wei F, Liu S, Song Y et al. TIARA: A visual exploratory text analytic system//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, 2010: 153-162

[24] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974, 19(6): 716-723

[25] Schwarz G. Estimating the dimension of a model. The Annals of Statistics, 1978, 6(2): 461-464

[26] Teh Y W, Jordan M I. Hierarchical Bayesian nonparametric models with applications//Hjort N, Holmes C, Müller P et al eds. Proceedings of the Bayesian Nonparametrics: Principles and Practice. UK: Cambridge University Press, 2009: 158-207

[27] Tang L, Liu H, Zhang J et al. Community evolution in dynamic multi-mode networks//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). New York, 2008: 677-685

[28] Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788-791

[29] Ding C, He X, Simon H. On the equivalence of nonnegative matrix factorization and spectral clustering//Proceedings of the SIAM Conference on Data Mining. Newport Beach, CA, USA, 2005: 606-610

[30] Ding C, Li T, Peng W et al. Orthogonal nonnegative matrix *t*-factorizations for clustering//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). New York, 2006: 126-135

[31] Wang L, Rege M, Dong M et al. Low-rank kernel matrix factorization for large scale evolutionary clustering. IEEE Transactions on Knowledge and Data Engineering, 2010, 24(6): 1036-1050

[32] Zhang Jianwen, Song Yangqiu, Chen Gang, Zhang Changshui. On-line evolutionary exponential family mixture//Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI). Pasadena, CA, USA, 2009: 1610-1615

[33] Jia Yangqing, Yan Shuicheng, Zhang Changshui. Semi-supervised classification on evolutionary data//Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI). Pasadena, CA, USA, 2009: 1083-1088

[34] Wang Y, Liu S X, Zhou L et al. Mining naturally smooth evolution of clusters from dynamic data//Proceedings of the SIAM Conference on Data Mining. Minneapolis, Minnesta, USA, 2007: 125-134

[35] Xu T, Zhang Z M, Yu P S et al. Evolutionary clustering by hierarchical Dirichlet process with hidden Markov state//Proceedings of the IEEE International Conference on Data Mining. Pisa, Italy, 2008: 658-667

[36] Xu T, Zhang Z M, Yu P S et al. Dirichlet process based evolutionary clustering//Proceedings of the IEEE International Conference on Data Mining. Pisa, Italy, 2008: 648-657

[37] Zhang Jianwen, Song Yangqiu, Zhang Changshui, Liu Shixia. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora//Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). Washington, DC, USA, 2010: 1079-1088

[38] Zhang Jianwen, Zhang Changshui. Multitask Bregman clustering//Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI). Atlanta, Georgia, USA, 2010: 655-660

[39] Zhang Jianwen, Zhang Changshui. Multitask Bregman clustering. Neurocomputing, 2011, 74(10): 1720-1734

[40] Murphy K P. Hidden semi-Markov models. Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA: Technical Report, 2002

[41] Yu S. Hidden semi-Markov models. Artificial Intelligence, 2010, 174(2): 215-243

[42] Blei D, Ng A, Jordan M et al. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3(4-5): 993-1022

[43] Wang C, Blei D, Heckerman D. Continuous time dynamic topic models//Proceedings of the Conference on Uncertainty in Artificial Intelligence. Helsinki, Finland, 2008: 579-586

ZHANG Chang-Shui, Ph. D., professor. His research interests include machine learning, pattern recognition, and the wide application areas of computer vision, music modeling, data mining, complex network, etc.

ZHANG Jian-Wen, Ph. D.. His research interests include machine learning and its applications on data mining.