多重选择决策树算法挖掘概念漂移数据流*

叶爱玲,刘 锋
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:重点研究了数据流分类挖掘中存在的概念漂移问题,并在CVFDT算法改进的基础上,提出了一种多重选择决策树算法mCVFDT。该算法将多重属性的选择机制加入到节点结构中,克服了CVFDT无法自动检测概念漂移的缺陷,同时避免了对决策树的重复遍历,提高了算法的分类精度和效率。实验结果证明该,算法随着样本数目的增加,在分类精度上比CVFDT算法有更好的表现。

关键词:数据流挖掘;多重选择;CVFDT;mCVFDT
中图分类号:TP311 **文献标识码**:A

Multiple-options decision tree mining concept-drifting data streams

YE Ai Ling, LIU Feng
(Dept. of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: This paper focuses on the concept-drifting data streams mining, and based on the CVFDT algorithm improvements proposed a multiple-choice decision-tree algorithm mCVFDT. In this algorithm multi-attribute selection mechanism is added to the node structure in an effort to overcome the CVFDT not automatically detect defects in the concept-drifting, while avoiding duplication of tree traversal algorithm to improve the classification accuracy and efficiency. Experimental results show that the algorithm increases with the number of examples in the classification accuracy than CVFDT algorithm has better performance.

Key words: data streams mining; multiple-options; CVFDT; mCVFDT

近年来,随着网络技术的普及和计算机技术的发展,海量的数据处理是各个领域面临的重要问题,如传感器网络、Web服务器、银行ATM终端、电子商务网站、股票交易、零售业终端等数据,都是一类新型的数据对象,具有大量、连续、快速、时变的特点,这些特点使得对流数据的分析和处理面临着巨大的挑战。因此成为当前数据挖掘领域研究的一个热点。

当前数据流分类最常用的是决策树模型,根据不同的分类需求选择最佳划分属性的统计度量也不同。数据流环境中,由于不可能重复扫描,因此,除了对响应时间和内存使用率方面有要求外,还需要解决概念漂移(concept drift)[1]问题。概念漂移是指目标分类模型随时间而改变的现象。在处理数据流分类时,概念漂移是需要重点考虑的问题。数据流分类算法中,一部分是对传统决策树的改进,以适应数据流的单次扫描需求,比如VFDT算法[2]和CVFDT算法[3];另一种方法是分类器系统(classifier ensemble)算法[4],该方法的思想是将流数据按照先后顺序分割成体积固定的数据块,然后利用集成学习(ensemble learning)方法,在每个分类器上训练基础分类器,最终组成集成分类器对流数据进行分类。这种方法虽然较好地解决了概念漂移,但数据块体积大小的选择会直接影响最终的分类精度。本文主要针对CVFDT算法进行改进,引入多重选择机制使CVFDT算法能够具备自我检测概念漂移的能力。

1 相关工作

DOMINGOS P等在文献[2]中提出增量决策树VFDT(Very Fast Decision Tree)算法,使用Hoeffding边界保证算法与批量学习的输出模型趋向一致。但该算法是在假定数据是从静态分布中随机获取的,不能反映数据随时间变化的特性。因此,HULTEN G等引入滑动窗口概念,提出了CVFDT(Concept adapting Very Fast Decision Tree)算法。CVFDT在某个节点的精度下降时生成一个备选子

* 基金项目:安徽省自然基金项目(070412051);安徽高校省级重点自然科学研究项目(KJ2007A43)

树，当备选子树的精度高于当前子树时，代替当前子树。

1.1 VFDT

VFDT 是一种基于 Hoeffding 边界针对数据流挖掘环境分类决策树的方法，它通过不断地将叶节点替换为分支节点而生成。该算法主要的创新在于利用 Hoeffding 不等式[5-6]确定叶节点变为分支节点所需要的样本数目。VFDT 中每个叶节点都保存相关属性值的统计信息，这些信息通过信息增益函数测试，当一个新样本到达，则自顶向下遍历决策树，到达树的叶节点，同时修改该叶节点上的统计信息。假设变量 r 取值范围为 R，观测 n 个样本后，样本观测平均值为 $\bar{r}$，则样本真值以置信度 $1-\delta$ 落在 $\bar{r}-\delta$ 区间范围内，其中

$$\varepsilon=\sqrt{\frac{R^2\ln(1/2)}{2n}} \tag{1}$$

VFDT 没有介绍相关连续属性的处理方法，因此在遇到连续属性的时候会造成很大程度的浪费。

1.2 CVFDT

CVFDT 算法是一种扩展的 VFDT 算法，它保持了 VFDT 的运行速度和分类精度，还具有处理样本产生过程中出现的概念漂移问题。CVFDT 通过维持一个样本的滑动窗口，在样本流入和流出窗体的时候更新已学习的决策树，使其与训练样本窗口保持一致，当 CVFDT 发现备选子树的精度远远大于原子树的时候，该子树将被替换。

然而，在 CVFDT 算法中备选子树并不是当前树的一部分，只有当备选子树的分类精度高于当前节点时才会替代当前子树。然而，可能存在使用备选子树的预测样本的类标签所得到的精度反而高于当前子树的情况。此外，如果旧的概念在备选子树替代完成后再次出现，CVFDT 则需要重新学习。CVFDT 算法的另一个缺陷在于无法自动检测概念漂移的发生，而只是通过不断地扫描树，比较精度来维持决策树的分类精度。

1.3 mCVFDT

mCVFDT(multiple Concept adapting Very Fast Decision Tree)算法采用多重选择机制，将当前主属性外的最佳预测精度属性和最近到达属性都加入到节点结构中，即将所有可能的子树都加入树结构本身，不存在备选子树。该方法避免了每个节点的划分只依赖于一个属性的情况，使得任何时候 mCVFDT 算法都能够保证有最新和最好的属性来用于分类预测。当旧的概念重新出现时，mCVFDT 可以立刻从自身的节点结构中已存在的备选属性中重新选择适合的分类模式，避免了对树的重复遍历。由于多属性都存在于节点结构中，即存在于整个决策树的结构中，因此可以更方便地使用备选属性做类标记的预测，从而提高分类精度。同时，在选择哪些可以作为模糊概念加入到节点结构的过程中，可将预测精度与当前节点分类精度作动态的比较，起到检测概念漂移的作用。

2 设计细节

2.1 节点结构

mCVFDT 将多种选择引入节点结构中，在 CVFDT 中，每个节点的划分只依赖于一个属性，其中关键是用哪个属性最适合充当这颗树的根节点。“信息增益”(Information Gain)[7]用来衡量一个属性区分以上数据样本的能力。信息增益量越大，该属性作为一棵树的根节点就更能使这棵树简洁。在 CVFDT 算法中，每个节点都是这样依赖一个属性进行划分，因此，它只能反映一个分类概念。如果每个节点都允许有多种选择，那么，该决策树就可以在不同的时段、不同的概念等情况下反映多重概念。

一个 mCVFDT 算法的节点结构如下：

mCVFDTnode={

X'：多选择的集合，该集合是 X 的一个子集。该节点依赖集合 X' 中的任何一个元素进行划分；X^i 是 X' 中的第 i 个元素；

n_{ijk}：统计信息集合，用于计算 X' 该节点属性 j 的取值为 i 的最终分类 k 个样本数目；n^l_{ijk} 对应 X' 中的第 l 个元素；

ST：开始时间，ST^i 记录每个 X^i 进入集合 X' 的时间；

Acc：预测精度，Acc^i 对应 X^i 的分类精度；

LT：预测精度提高的最近实时时间；LT^i 对应 X^i 的预测精度提高所用的最近实时时间；

}

X' 为每个节点提供了多种选择。开始时 X' 为空，当样本到达后，如果 Hoeffding bound 测试显示该属性达到用作划分属性的条件，则将其加入到 X' 中，同时更新 n_{ijk} 和 ST，从此时起 Acc 和 LT 也随着样本的到达而更新，一个 mCVFDTnode 子节点对应于每个该属性值随之创建。当算法检测出某个属性可能有更好的分类效果，而并不在集合 X' 中时，则将其加入到 X' 中。

ST 和 Acc 可以用作预测类标签。可以通过计算开始时间和预测精度来决定选择 X' 中的哪个元素作为划分属性。因此，用户在使用该学习模型来进行分类预测时，既可以选择最近的元素，也可以选择预测精度最高的元素，还可以根据需要采取加权方式考虑这两种方式。

如果不断有新的属性加入到集合 X' 中来，那么该集合就会变得非常巨大，因此需要定期删除一定的属性，此时需要查看 LT，LT 中记录的是 X' 中每个属性最近一次的预测精度提高的时间，用 LT 可以检测出哪些属性的预测精度在相当长一段时间内没有提高，便可以认为这些属性不适合描述多数当前的概念，因此先将其移除。

2.2 概念漂移检测机制

设当前的学习模式为 MT(mCVFDT)，当一个样本成功到达，MT 首先预测该样本的类标签，此时，MT 的预测精度就相应地发生改变，这种变化的幅度可以用作对概

念漂移的检测。可以用贝叶斯分类器来进行检测。

设 MT 的预测精度为 p_0，让 MT 对 m 个到达的样本进行类标签的预测，此时预测精度满足二项分布 $b(m,p_0)$。另假设 H_0 表示 MT 树没有变化，即当前平均预测精度 p 与 p_0 相比没有降低。那么，对立假设 H_1 表示平均预测精度有明显降低，此时 $p<p_0$。决策规则为，若 m 个样本的平均预测精度 p 发生了明显的变化 $a=0.05$，则舍弃 H_0，接受 H_1。在这种情况下，由于在 H_0 时，$u(=mp)$ 是 $N(mp, mp_0(1-mp_0))$ 满足二项分布的近似值，因此临界区域可以用如下公式进行计算：

$$\frac{u-u_0}{\sigma/\sqrt{m}} \leqslant -z(a) \tag{2}$$

其中 $u_0=mp_0$，$\sigma=mp_0(1-p_0)$。等式(2)称为标准化检验统计量。若发现 $z(u-u_0)/(\sigma/\sqrt{m})$ 小于 $-z(a)=-z(0.05)=-1.645$，则舍弃 H_0，接受 H_1：$u<u_0$。即，此时发生了概念漂移；反之，如果 $z>-z(a)$，则不舍弃 H_0，即当前没有发生概念漂移。

2.3 mCVFDT 算法的类标记预测过程

mCVFDT 算法的预测方法过程为：

输入：StartNode 为根节点

X：一个未标记的样本

输出：预测过的类标签 X

则执行：

(1)将 StartNode 传给 CurrentNode；

(2)若 CurrentNode 非空；

(3)if CurrentNode(X')是个空集合；

(4)将 y 标记为 CurrentNode 的多数类；

(5)将 CurrentNode 置空；

(6)else ；

(7)将 CurrentNode(X'^{best})标记为 CurrentNode (X')中预测精度最高的元素；

(8)将子节点依照 X 中的 CurrentNode(X')的值加入到 CurrentNode。

3 实验结果分析

本文设计的实验主要是比较 mCVFDT 和 CVFDT 在均衡数据流环境下的分类精度，并未特别区分是离散属性还是连续属性。实验环境是：奔腾 IV/2 G 的 CPU，内存 512 MB，操作系统为 Windows XP。实验数据采用超平面[8]模拟的数据流。系统所用参数为：$\delta=0.000\ 1$，$\tau=0.05$，$\omega=1\ 000\ 000$，$n_{min}=300$。

一个 d 维的超平面可描述为：

$$\sum_{i=1}^{d} a_i x_i = a_0 \tag{3}$$

其中权值 $a_i(1\leqslant i\leqslant d)$ 用[0,1]之间的随机数来赋值。

当 $d=10$ 时，mCVFDT 和 CVFDT 在不同的概念漂移变化水平参数 $k=2,4,6$ 和 8 时的不同精度表现，结果如图 1 所示。

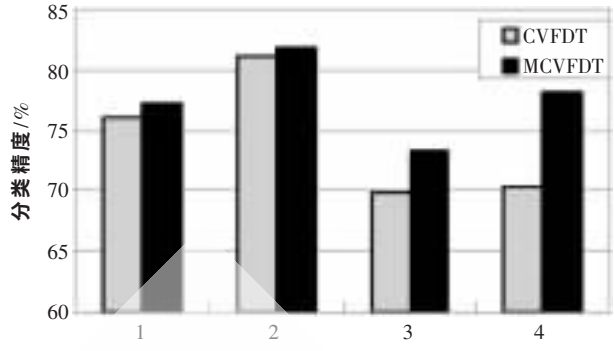

图 1 不同概念漂移水平下的分类精度

再比较当确定概念漂移参数 $k=2,4,6$ 和 8 时，在不同数量的样本时两种算法的分类精度表现。图 2 表示当 $k=8$ 时，随着样本数的增加，算法的分类精度的变化。

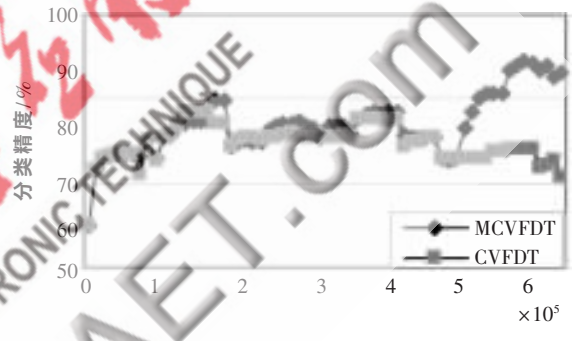

图 2 相同概念漂移水平下不同样本数的分类精度

从图 2 中可以看出，mCVFDT 算法在样本数小于 500 000 时与 CVFDT 有相似的表现，但当样本数高于 500 000 时分类精度有了明显的差异。因此，在样本数目庞大的环境中 mCVFDT 比 CVFDT 有更好的表现。

本文主要从节点结构上对 CVFDT 算法进行了改进，这种方法可以明显提高算法的分类精度和对概念漂移的处理能力，大大提高了算法的整体效率。但是本文实验所使用的数据是模拟数据流，在实际中的应用还有待进一步研究。

参考文献

[1] JEFFREY C S, RICHARD H G. Beyond incremental processing[J]. Tracking Concept Drift Proceedings of the Fifth National Conference on Artificial Intelligence, 1986.

[2] ROWL C H. Covert channels in the TCP/IP protocol suite [J]. Tech. Rep. 5, First Monday, Peer Reviewed Journal on the Internet, 1997(7).

[3] HULEN G, SPENCER L, DOMINGOS P. Mining time changing data streams[C]. Proc. of the 6th ACM SIGKDD Int. Conf. On Knowledge Discovery and Data Mining. San Francisco, 2001.

[4] WANG H W, FAN P. YU. Mining concept drifting data streams using ensemble classifiers[C]. The 9th ACM Int

Conf.on Knowledge Discovery and Data Mining.Washington: ACM Press, 2003: 228-235.

[5] HOEFFDING W.Probability inequalities for sums of bounded random variables[J].American Statistical Association, 1963(5): 14-30.

[6] MARON O.A moore1 hoeffding races: accelerating model selection search for classification and function approximation [J].Advances in Neural Information Processing Systems, 1993(6): 59-67.

[7] LI X, BARAJAS J M, DING Y.Collaborative filtering on streaming data with interest-drifting[J].Intell.Data Anal. 2007, 11(1): 75-87.

[8] JIANG J.A literature survey on domain adaptation of statistical classifiers[EB/OL].[2009-11-10].http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey.2008.

(收稿日期: 2009-11-19)

作者简介:

叶爱玲, 女, 1983年生, 硕士研究生, 主要研究方向: 数据流挖掘。

刘锋, 男, 1962年生, 教授, 主要研究方向: 软件工程和并行计算。