ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 2,Mar. 2016,pp. 315－324
DOI:10.16337/j.1004-9037.2016.02.011

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于主要特征抽取的重现概念漂移处理算法

冯　超[1]　文益民[1,2]　汤凌冰[3]

(1.桂林电子科技大学计算机科学与工程学院,桂林,541004;2.桂林电子科技大学广西可信软件重点实验室,桂林,541004;3.湖南商学院计信学院,长沙,410205)

摘　要:针对重现概念漂移检测中的概念表征和分类器选择问题,提出了一种适用于含重现概念漂移的数据流分类的算法——基于主要特征抽取的概念聚类和预测算法(Conceptual clustering and prediction through main feature extraction, MFCCP)。MFCCP通过计算不同批次样本的主要特征及影响因子的差异度以识别重复出现的概念,为每个概念维持且及时更新一个分类器,并依据Hoeffding不等式选择最合适的分类器对当前样本集实施分类,以提高对概念漂移的反应能力。在3个数据集上的实验表明:MFCCP在含重现概念漂移的数据集上的分类准确率,对概念漂移的反应能力及对概念漂移检测的准确率均明显优于其他4种对比算法,且MFCCP也适用于对不含重现概念漂移的数据流进行分类。

关键词:重现概念漂移;主要特征;影响因子;数据流;Hoeffding不等式

中图分类号:TP181　　**文献标志码**:A

Algorithm of Recurring Concept Drift Based on Main Feature Extraction

Feng Chao[1], Wen Yimin[1,2], Tang Lingbing[3]

(1. School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin, 541004, China; 2. Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, 541004, China; 3. Computer and Information Engineering School, Hunan University of Commerce, Changsha, 410205, China)

Abstract: Recurring concept drift is one of the sub-types of concept drift. In recurring concept drift detection, it is very important to represent concepts and select the most appropriate classifier to classify. We propose an algorithm, conceptual clustering and prediction through main feature extraction (MFCCP), for classifying data stream with recurring concept drifts. MFCCP can recognize recurring concepts by computing the differences of main features and impact factors of different batches of samples. It maintains a classifier for each concept and monitors the classification accuracy to select classifier according to hoeffding inequality in order to enhance the ability of adapting to concept drift. The experimental results over the three datasets illustrate that MFCCP achieves better classification accuracy, adapts faster to concept drift, and detects concept drift more accurately than the other four algorithms on the data streams with recurring concept drifts, and therefore, MFCCP is apt to classify data stream without recurring concept drift.

Key words: recurring concept drift; main feature; impact factor; data stream; hoeffding inequality

---

基金项目:国家自然科学基金(61363029,61540053)资助项目;广西区科学研究与技术开发(桂科攻14124005-2-1)资助项目;湖南省博士后科研专项资助计划(2011RS4073)资助项目;广西信息科学中心(YB408)资助项目。
收稿日期:2015-05-20;**修订日期**:2015-07-21

引 言

“概念漂移”一词由 Schlimmer 和 Granger 在 1986 年首次提出，指在不同时间段生成数据的分布发生了变化[1]。在对数据流进行分类时，由于数据流的持续时间长，容易产生概念漂移，使得分类器必须动态调整。针对数据流中的概念漂移问题，国内外学者开展了一系列的研究工作，文献[2～9]从各自角度综述了国内外学者对概念漂移的研究。“重现概念漂移”是概念漂移的一种，指之前出现过的概念在将来可能会再次出现，但出现的时间不能确定[10]。如在垃圾邮件分类问题中，某一用户在不同时间段对垃圾邮件的定义可能会反复变化，对于某位正在求职的用户，关于招聘的邮件会被视为正常邮件，而当此用户获得某岗位后，就会把这种邮件当成垃圾邮件，而当以后该用户想更换工作时，又会重新对此种主题的邮件感兴趣，而将其又定义为正常邮件. 又如可穿戴设备检测到的各种人体状态也常常会重复出现。

尽管在实际中概念重复出现[10]很普遍，但在已有处理概念漂移的分类算法却很少考虑到在不同时间段可能会有重复出现的概念。比如，批量学习法[11]、滑动窗口法[12]和样本加权法[13]都直接丢弃了旧的分类器。这在处理含重现概念漂移的数据流时会导致对重现概念的适应迟滞，分类准确率降低。在处理重现概念漂移时主要面临以下几大难题：(1)当多个概念重复出现时，如果没有分类器选择机制，即使有更准确的概念漂移检测机制也难以保证较高的分类准确率；(2)判定重现概念漂移时，必须以某种形式表征概念并存储，表征概念的难点在于既需满足概念表征形式的存储容量为常量，又需尽可能反映出数据分布。

研究者们提出了不同的处理重现概念漂移算法[10,14-20]。文献[14]用正样本和负样本的每维特征取不同值的数量表征样本所属概念；文献[19]提出的元学习器算法(Meta-learner，META)和文献[20]提出的重现概念漂移框架(Recurring concept drift framework，RCD)只用样本的特征集(不考虑类别)表征样本所属概念；文献[10]提出的概念预测算法(Conceptual clustering and prediction，CPP)使用样本类别标签与每一维特征之间的概率关系表征样本所属概念；文献[15～18]的算法则都利用样本训练的分类器表征样本所属概念。

CCP 算法[10]的主要创新之处在于用概念向量(Conceptual vector，CV)表达和存储多个概念。概念向量由样本特征与类别间的概率关系组成。通过将数据块转化为概念向量，计算不同数据块对应的概念向量之间的距离，并与设定的阈值比较以判断是否属于相同概念。文献[10]提出了一种适合于重现概念漂移问题的分类框架 CCP，主要包括 3 部分：(1)转换函数。用于将一批样本转换为概念向量；(2)增量式聚类算法。用于将相似的概念向量聚成一个簇；(3)增量式分类器。它实现对属于同一概念的样本的增量学习。文献[10]提出的概念向量非常适用于重现概念漂移问题，但是该算法在某些领域如邮件分类时存在一些不足：(1)概念漂移的检测容易受到噪声的影响。当样本特征的维数很高时，噪声所造成的差别会占据较大的比重，容易混淆噪声造成的差别与概念漂移造成的差别。(2)将不影响样本类别的特征的差别也考虑在内。事实上，对于邮件分类问题，在某个时间段其实只有少数特征会影响样本的类别，将不影响样本类别的特征考虑在内会给概念漂移检测造成困难。另外，CCP 算法每隔固定窗口的样本进行重现概念的检测并更新当前分类器。概念的实质是全部样本的分布，需要较大的样本量才能较准确地提取所属概念，因此较小的窗口会导致概念的表征不准确，从而造成样本分类不准确，而较大的窗口会导致对概念漂移反应迟钝，尤其在多个概念变化频繁的情形下，会造成样本分类准确率大幅降低。

针对以上不足，本文提出了一种基于主要特征抽取的重现概念漂移处理算法(Conceptual clustering and prediction through main feature extraction，MFCCP)。该算法从样本所有的特征维中选取对类别影响最大的一部分特征，称作主要特征，利用主要特征及其对样本类别的影响来计算不同样本集所对应概

念向量之间的距离,以检测是否发生了概念漂移。另外,为了解决数据窗口的大小问题,本文提出设置两个窗口,每隔较大窗口计算概念向量并进行重现概念漂移的检测,每隔较小窗口根据当前分类器的准确率选择分类器实施分类,从而提高分类器对概念漂移的适应性。

1　MFCCP 算法

MFCCP 算法主要适用于特征为二值类型的两类分类问题,样本的特征及类别只有 1 或 0 两种取值。对于邮件分类问题,第 i 维特征值为 1 表示第 i 维特征对应的词出现在此封邮件中,为 0 则表示不存在。类别为 1 表示正常邮件,为 0 表示垃圾邮件。假设样本特征维数为 n,本文定义一种新的概念向量 $\boldsymbol{Z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n,m)$,其中 m 表示此概念向量由 m 个样本转化而来,$\boldsymbol{z}_i(i=1,2,\cdots,n)$为一个三维向量$(z_{i1},z_{i2},z_{i3})$,其中 z_{i1} 表示这 m 个样本中第 i 维特征值为 1 且类别为 1 的样本个数;z_{i2} 表示这 m 个样本中第 i 维特征值为 1 且类别为 0 的样本个数;z_{i3} 表示第 i 维特征值为 1 的样本个数。

为了计算不同样本集对应的概念向量之间的距离,本文引入了主要特征及其影响因子。主要特征为对样本类别有较大影响的特征,特征对类别的影响用影响因子(Impact factor, IF)表示。IF[i]表示第 i 维特征值为 1 时样本类别为 1 与 0 的概率的差距。其计算式为

$$\begin{aligned}\mathrm{IF}[i]&=P(y=1\mid x_i=1)-P(y=0\mid x_i=1)=\\&(P(y=1,x_i=1)-P(y=0,x_i=1))/P(x_i=1)=(z_{i1}-z_{i2})/z_{i3}\end{aligned}\tag{1}$$

在邮件分类中即为由于某一个特征词的出现,此邮件是正常邮件还是垃圾邮件的概率的差别。选取最小的 k 值,使得前 k 个特征的影响因子绝对值之和,占所有特征的影响因子绝对值之和的 98%以上,则此 k 个特征即为该概念向量对应的主要特征。

MFCCP 的核心在于概念向量的增量聚类。聚类过程主要涉及 3 个算法:(1)convertToConceptVec 算法用于计算一个样本集对应的概念向量;(2)findSameConcept 算法用于查找与待聚类向量属同一概念的向量;(3)addConceptVec 算法用于合并两个属同一概念的概念向量。

MFCCP 利用算法 convertToConceptVec 计算当前样本集所对应的概念向量。利用算法 findSameConcept 计算待聚类概念向量与已保存概念向量集合中各概念向量的差异度,若最小差异度小于阈值则认为当前样本集属于重现概念,利用算法 addConceptVec 更新与其差异度最小的概念向量;否则认为当前样本集属于新概念,将其概念向量保存至概念向量集合中。

算法 1　convertToConceptVec()

输入　一个样本集

输出　该样本集对应的概念向量

(1)计算样本集中样本数量 m,特征维数 n;

(2)for $i=1$ to n,依次执行步骤:(a)计算第 i 维特征值为 1 且类别为 1 的样本数量 z_{i1};(b)计算第 i 维特征值为 1 且类别为 0 的样本个数量 z_{i2};(c)计算第 i 维特征值为 1 的样本数量 z_{i3}。

(3)返回概念向量 $\boldsymbol{Z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n,m)$。

算法 2　findSameConcept()

输入　已保存的概念向量集合 CVS;一个待聚类的概念向量 curCV;阈值 $\theta\in[0,1]$。

输出　CVS 中与 curCV 属同一概念的元素索引 index 或-1(-1 表示 CVS 中不存在与 curCV 属同一概念的概念向量)。

(1)对于已保存的 curCV 及 CVS 中的每个概念向量,依次执行步骤(a～e),以计算每个概念向量对应的主要特征及主要特征的影响因子:

(a)将出现的样本数量少于此概念向量对应的样本数量的 1%的特征的影响因子设为 0;(b)计算余下每一维特征的影响因子 IF[i] $=(z_{i1}-z_{i2})/z_{i3}$;(c)根据各特征的影响因子的绝对值从大到小排序,得

到排序后的特征集合;(d)选取最小 k 值使得前 k 个特征的影响因子绝对值之和占所有特征的影响因子绝对值之和的 98%以上;(e)选择前 k 个特征作为此概念向量对应的主要特征,保存主要特征及其影响因子。

(2)对于 CVS 每一个概念向量,依次计算其和 curCV 的主要特征个数之和,及这两个概念向量对应的主要特征的交集中特征的数量,分别记作 TN 和 CN,定义 DN=TN−2×CN。

$$\mathrm{DF}_1 = \frac{\mathrm{DN}}{(\mathrm{TN}-\mathrm{CN})} \tag{2}$$

$$\mathrm{DF}_2 = \sum_{i\in \text{交集}} (|\ \mathrm{IF}_1[i]-\mathrm{IF}_2[i]\ |) / (2\times \mathrm{CN}) \tag{3}$$

$$\mathrm{DF} = \alpha\times \mathrm{DF}_1 + (1-\alpha)\times \mathrm{DF}_2 \tag{4}$$

式中:α, DF_1, DF_2, $\mathrm{DF}\in[0,1]$。

(3)计算所有得到的 DF 中的最小值 $\mathrm{DF}_{\min}$ 及对应的概念向量在 CVS 中的索引 index,判断 $\mathrm{DF}_{\min}$ 与阈值 θ 的大小。若小于 θ,则输出 index;否则,输出−1。

算法 3 addConceptVec ( )

输入 两个概念向量。

输出 合并后的概念向量。

(1)将两个概念向量的每一对应分量相加即为合并后的概念向量。

(2)返回合并后的概念向量。

本文算法 MFCCP 步骤描述如下:

(1)初始化概念向量集合和分类器集合为空,然后转到步骤(2)。概念向量集合用于保存各个概念对应的概念向量,分类器集合用于保存各个概念对应的分类器,对于每个概念都会保存其对应的一个概念向量和一个分类器。

(2)读取输入样本。每到达一个样本采用当前分类器进行分类,若输入样本为空则算法结束。当前分类器即从分类器集合中挑选的用于分类当前样本的分类器,其分类结果被作为最终分类结果以评价算法的性能。在分类第 1 个窗口的样本时由于尚无分类器,因此通过随机猜测分类。若到达样本个数为 sw 的整数倍,则转到步骤(3),sw 取窗口大小 w 的 $1/N(N=2,3,4\cdots)$倍。

(3)每到达 sw 个样本调整当前分类器,执行完后,判断到达样本个数是否为窗口大小的整数倍,若是则转到步骤(4);否则转到步骤(2)。具体细节如下:每分类 sw 个样本后获得这些样本的真实类别,计算当前分类器对这些样本的分类准确率 acc。对于每个分类器都会保存对应的历史准确率 histAcc,即分类器对与其同属 1 个概念的样本的分类准确率。用 curHistAcc 表示当前分类器的历史准确率,若 curHistAcc−acc>ε,则用分类器集合中的各分类器分别对此 sw 个样本分类,并选取分类准确率最高的分类器作为当前分类器,ε 的值通过 Hoeffding 不等式确定[21]。

Hoeffding 不等式:假设随机变量 $X_1,\cdots,X_n$ 服从参数为 β 的伯努利分布 Bernoulli(β),则对任意 $\varepsilon>0$,有

$$P(\beta-\overline{X}_n>\varepsilon)\leqslant \mathrm{e}^{-2n\varepsilon^2} \tag{5}$$

式中:$\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$。

分类器预测结果正确与否服从伯努利分布[22]。若假设当前样本集与用来分类它的分类器属同一概念,显著性水平取 0.05。要使该假设成立,则允许的 curHistAccc 与 acc 的最大差值为

$$\varepsilon = \sqrt{\ln 0.05/(-2n)} \tag{6}$$

式中:n 取值为 sw;curHistAcc 相当于式(5)中的 β;acc 相当于式(5)中的 $\overline{X}_n$。

(4)每到达整个窗口的样本进行重现概念漂移检测,并根据检测结果更新概念向量集合及分类器集

合,调整当前分类器,执行完毕转入步骤(2)。具体细节如下:首先利用 convertToConceptVec 算法将当前窗口样本转换为概念向量,用 curConceptVec 表示;然后通过 findSameConcept 算法从已保存的概念向量集合中搜索与 curConceptVec 属同一概念的概念向量。若 findSameConcept 返回结果 index 为-1,表示当前窗口样本属于新概念,则将 curConceptVec 加入概念向量集合,并利用当前窗口样本训练分类器,将此分类器的历史准确率 histAcc 和曾经分类过的与其同属一个概念的样本数量 p_n 赋值为 0,然后加入分类器集合,且将此分类器作为当前分类器;若 index 不等于-1,表示当前窗口样本属于重现概念,则通过 addConceptVec 算法利用 curConceptVec 更新概念向量集合中索引为 index 的概念向量,用 targetClf(targetClf 表示分类器集合中索引为 index 的分类器)分类当前窗口样本得到分类准确率 acc,更新 targetClf 的历史准确率为

$$\text{newHistAcc}=p_n/(p_n+w)\times\text{oldHistAcc}+w/(p_n+w)\times\text{acc}$$

式中:oldHistAcc 为 targetClf 原历史准确率;w 为窗口大小,更新 targetClf 的 p_n 为原 p_n 加上 w,用当前窗口样本更新 targetClf,并将其设为当前分类器。

2　实验与分析

2.1　评价标准

为了评价 MFCCP 算法的有效性,本文采用了 4 种评价标准:(1)准确率(Accuracy)。所有正确分类的样本数量占总样本数量的比例。(2)精确率(Precision)。所有正确分类的正例样本数量占分类为正例的样本数量的比例。(3)召回率(Recall)。所有正确分类的正例样本数量占总正例样本数量的比例。(4)耗时(Time)。算法对一个数据集进行学习和分类的总时间,以 s 为单位。

2.2　实验数据

本文实验中使用了 3 个数据集。

第 1 个数据集为 elist dataset[10],包含 1 500 个邮件样本,特征维数为 913。该数据集模拟了同一用户在不同时间段对某一主题兴趣的反复突然变化,包括医药、太空及棒球 3 个主题。其中第 1～300,601～900,1 201～1 500 个样本属于同一概念,表示用户对医药的主题感兴趣,而对太空和棒球的主题不感兴趣;第 301～600,901～1 200 个样本表示用户对太空和棒球的主题感兴趣,而对医药的主题不感兴趣。

第 2 个数据集是按照文献[23]中的数据生成方式生成的 SEA dataset。共生成 3 600 个样本,特征维数为 10,每维特征值为 0 或 1,共 3 个概念,数据包含了 10%的噪音。当 $f_1+f_2+\cdots+f_7<\gamma$($f_i$ 表示样本的第 i 维特征)时类别为 1,否则类别为 0。其中第 1～600,1 801～2 400 个样本属于概念 1:$\gamma=1$;第 601～1 200,2 401～3 000 个样本属于概念 2:$\gamma=5$;第 1 201～1 800,3 001～3 600 个样本属于概念 3:$\gamma=7$。

第 3 个数据集 spam dataset 代表了不含重现概念漂移的数据,为文献[10]从 Spam Assassin Collection 提取出的邮件数据集,其中含 9 300 个样本,有 499 维属性。

2.3　对比算法

动态加权多数投票算法(Dynamic weighted majority,DWM)[24]:保存 1 个由多个分类器组成的集成分类器且对每个分类器都赋以权值,集成分类器分类时样本的类别通过所有分类器加权投票决定。当单个分类器分类错误时就减小其权值,当该权值小于预设的阈值时就将相应的分类器从集成分类器中移除;当集成分类器分类错误时新的分类器就会以权值 1 被加入集成分类器中。当获取样本的真实类别后,用来更新所有的分类器。本文实验中参数按照文献[24]设置($\beta=0.5$,$\theta=0.01$,$p=1$),基分类器为 Naive bayes。

META 是一种基于元学习器的算法。算法中的分类器分为两层，Level0 层的 Normal Classifier 和 Level1 层的 Referee。Level0 层分类器用来预测样本标签，如果 Level0 分类正确，则 Level1 样本标签为“+”，否则为“−”。Level1 层分类器用来选择最合适的 Level0 层分类器来对下一批样本做分类。算法保存了所有已出现概念的 Level0 和 Level1 层分类器，当原有的概念重复出现时，已存储的分类器便能被再次利用。本文实验中参数按照文献[19]设置($\theta=0.7$)，基分类器为 Naive bayes。

RCD 算法保存每个概念对应的分类器及其训练样本，通过概念漂移检测方法(Drift detection method，DDM)[25]判断是否发生概念漂移，当 DDM 检测到概念漂移后利用非参统计方法判断当前概念是否属于原有概念。若属于，则更新对应的分类器；若为新出现的概念，则保存下来并为之训练一个分类器。本文实验中参数按照文献[20]设置(即 $s=0.01$，$b=200$，$t=200$，$k=5$)，基分类器为结合朴素贝叶斯的 Heffding 树(Heffding tree with Naive Bayes，HTNB)。

CCP 算法在分类过程中当一定数量样本的类别获知后，该算法利用转换函数将该批样本映射成新的概念表示模型，对每个概念训练一个分类器，并利用聚类算法判断最近获得的一批样本是否属于已有的概念。若属于，则更新对应的已有概念及分类器；若为新出现的概念，则保存下来并为之训练一个分类器。本文实验中参数按照文献[10]设置(即 3 个数据集中分别为 $\theta=4$，$\theta=2.5$，$\theta=0.6$)，基分类器为 Naive bayes。

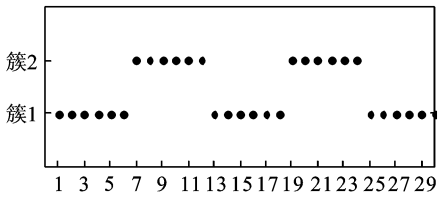
MFCCP 在 elist dataset 上的参数设置为 $w=50$，$sw=25$，$\alpha=0.3$，$\theta=0.6$。其余两个数据集上的参数设置为 $w=100$，$sw=20$，$\alpha=0.3$，$\theta=0.6$，基分类器为 Naive bayes。

2.4 实验结果与分析

为了对比本文与文献[10]提出的概念表征方式，图 1 描述了 MFCCP 与 CCP 在 elist dataset 和 SEA dataset 上的概念向量增量聚类结果。图 1(a～h)中横坐标表示概念向量序号，纵坐标表示概念向量聚类后的簇，纵坐标相同的概念向量属于同一个簇，w 表示窗口的大小。从图 1 中可以看出：MFCCP 在不同大小的窗口条件下和两个不同数据集上都能完全正确地区分出属于不同概念的向量，将属于相同概念的概念向量聚到同一个簇中；而 CCP 在这两个数据集上对概念向量的聚类结果都不理想，或是将属于同一概念的向量聚到不同簇，或是将属于不同概念的向量聚到同一个簇。对比结果说明，通过抽取主要特征的方式能够更准确地区分样本集所属的概念。

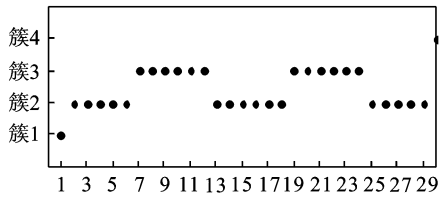
为了分析 MFCCP 的分类性能，表 1～3 对比了其和另外 4 种算法分别在 elist dataset，SEA dataset 和 spam dataset 上的分类准确率。为了降低实验结果的随机性，本文将 elist dataset 按样本的先后顺序划分为 5 个数据块，将属同一概念的各个数据块之间的所有排序进行组合后生成 12 份数据集，概念出现的顺序仍保持不变，实验重复 12 次。另外，本文随机生成了 100 份 SEA 数据集，概念出现的顺序仍保持不变，实验重复 100 次。从表 1～3 可以看出：在 elist 数据集和 SEA 数据集上，除运行时间外，MFCCP 明显优于其他 4 种算法；在 spam 数据集上，MFCCP 在准确率和精确率评价指标上表现最好。相比于 CCP，MFCCP 在前 3 个指标上均表现更好，由于其算法更复杂，所以运行时间稍微长一点。从这 3 个表中还可以看到：由于 DWM 没有考虑概念的重复出现，因此在含重现概念漂移的数据集 elist 和 SEA 上分类精度都较低，这说明了检测并利用重现概念十分必要。META 虽然考虑了重复出现的概念，但它在检测重现概念时只考虑了特征之间的相似性，没有考虑样本类别，使其不适用于含分界面变化较大而特征分布几乎不变的概念漂移的数据集，而 elist dataset 正是这种类型的数据，因此 META 在 elist dataset 上分类精度较差。RCD 在 3 个数据上表现都较差，是因为 RCD 在比较两批样本是否属于同一概念时只考虑了特征集的分布 $P(X)$ 的相似度，导致在发生实概念漂移时效果较差。

图 2～4 分别表示 DWM，META，CCP，RCD 和 MFCCP 5 种算法在 elist dataset，SEA dataset 以及



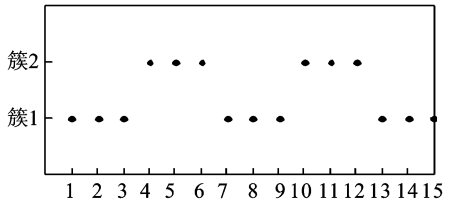
(a) w=50时MFCCP在elist dataset上的概念向量聚类结果

(a) CV clustering result of MFCCP over elist dataset with w=50



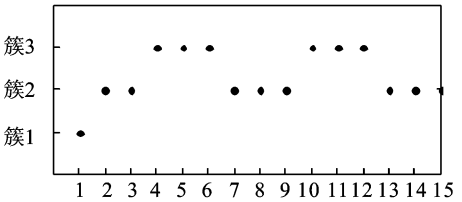
(b) w=50时CCP在elist dataset上的概念向量聚类结果

(b) CV clustering result of CCP over elist dataset with w=50



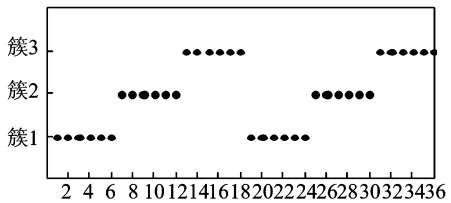
(c) w=100时MFCCP在elist dataset上的概念向量聚类结果

(c) CV clustering result of MFCCP over elist dataset with w=100



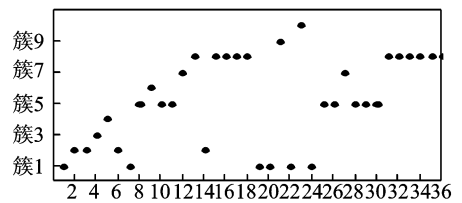
(d) w=100时CCP在elist dataset上的概念向量聚类结果

(d) CV clustering result of CCP over elist dataset with w=100



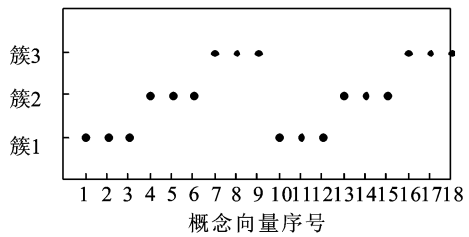
(e) w=100时MFCCP在SEA dataset上的概念向量聚类结果

(e) CV clustering result of MFCCP over SEA dataset with w=100



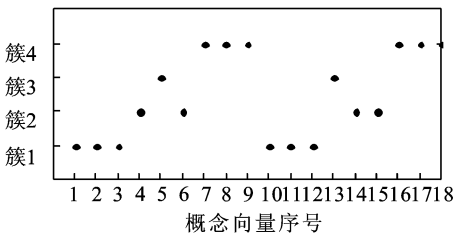
(f) w=100时CCP在SEA dataset上的概念向量聚类结果

(f) CV clustering result of CCP over SEA dataset with w=100



(g) w=200时MFCCP在SEA dataset上的概念向量聚类结果

(g) CV clustering result of MFCCP over SEA dataset with w=200



(h) w=200时CCP在SEA dataset上的概念向量聚类结果

(h) CV clustering result of CCP over SEA dataset with w=200

图 1　MFCCP 和 CCP 在 elist dataset 和 SEA dataset 上的概念向量聚类结果

Fig. 1　Concept vector clustering results of MFCCP and CCP on elist dataset and SEA dataset

spam dataset 上随着样本到来分类准确率的变化。图 2 中概念漂移发生在第 300,600 和 1 200 个样本处,其中第 600,1 200 发生重现概念漂移。从图 2 可以看出,在 elist dataset 上 DWM 的表现最差,准确率始终低于 0.5;在概念漂移发生后 MFCCP 和 CCP 的分类准确率能够迅速恢复到正常水平,而 META 和 RCD 需要较长时间。这是由于 MFCCP 和 CCP 在 elist dataset 上的概念漂移检测比 META 和 RCD 更准确。图 3 中概念漂移发生在第 600,1 200,1 800,2 400 和 3 000 个样本处,其中第 1 800,2 400 和 3 000处发生重现概念漂移。从图 3 可以看出,MFCCP 和 META 表现最好,而且准确率波动小于其他 3 种算法,RCD 和 DWM 在发生概念漂移后的恢复速度明显慢于其他 3 种算法。

表 1 DWM, META, RCD, CCP 和 MFCCP 在 elist dataset 上的性能

Tab. 1 Performance of DWM, META, RCD, CCP and MFCCP over the elist dataset

算法	准确率	精确率	召回率	耗时/s
DWM	0.420±0.004	0.456±0.001	0.460±0.001	12.068±0.551
META	0.569±0.018	0.613±0.008	0.523±0.091	10.552±0.090
RCD	0.623±0.034	0.656±0.040	0.610±0.030	12.117±0.219
CCP (w=50)	0.749±0.085	0.773±0.086	0.748±0.071	4.521±0.107
MFCCP (w=50)	0.847±0.011	0.859±0.013	0.849±0.016	5.404±0.197

表 2 DWM, META, RCD, CCP 和 MFCCP 在 SEA dataset 上的性能

Tab. 2 Performance of DWM, META, RCD, CCP and MFCCP over the SEA dataset

算法	准确率	精确率	召回率	耗时/s
DWM	0. 687±0.011	0.695±0.011	0.638±0.011	0.278±0.013
META	0.787±0.012	0.792±0.015	0.764±0.014	0.385±0.023
RCD	0.701±0.022	0.716±0.026	0.620±0.019	0.330±0.030
CCP(w=100)	0.764±0.030	0.752±0.041	0.754±0.030	0.196±0.027
MFCCP(w=100)	0.805±0.013	0.796±0.018	0.794±0.017	0.228±0.024

表 3 DWM, META, RCD, CCP 和 MFCCP 在 spam dataset 上的性能

Tab. 3 Performance of DWM, META, RCD, CCP and MFCCP over the spam dataset

算法	准确率	精确率	召回率	耗时/s
DWM	0.917	0.845	0.827	6.120
META	0.911	0.808	0.857	11.395
RCD	0.898	0.806	0.792	14.706
CCP(w=100)	0.906	0.828	0.784	5.652
MFCCP(w=100)	0.918	0.852	0.807	7.001

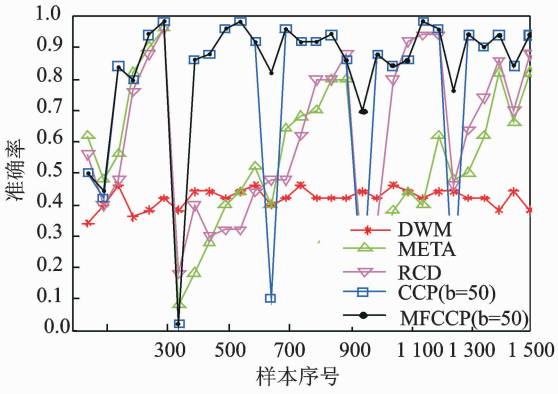


图 2 在 elist dataset 上的准确率

Fig. 2 Accuracy over elist dataset

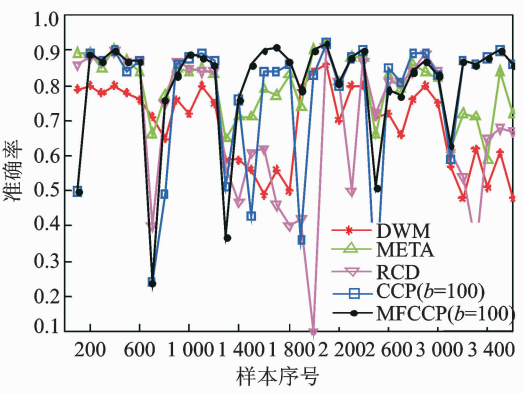


图 3 在 SEA dataset 上的准确率

Fig 3 Accuracy over SEA dataset

图 4 表示 5 种算法在 spam dataset 上每 100 个样本的分类准确率，概念漂移点未知。从图 4 可以看到，在 spam dataset 上 5 种算法表现比较相近。

从图 2,3 还可以看出，重现概念漂移发生时，MFCCP 的准确率下降幅度远小于 CCP。这是由于 MFCCP 不断监测当前分类器的准确率。当概念漂移发生时，MFCCP 的准确率突然降低，MFCCP 会根据

Hoeffding 不等式判断是否应该更换当前分类器,及时从已保存的分类器中选择最合适的分类器对最近批次的样本分类。所以每当发生重现概念漂移时,MFCCP 能选择已保存的对应于重现概念的分类器对最近获得的样本进行分类。另外,从图 3 中还可以看出,MFCCP 只有在概念漂移发生时才出现准确率的大幅下降,而 CCP 在正常时期也出现准确率的大幅下降,这是由于 CCP 误判概念漂移(见图 1)造成,这进一步说明了 MFCCP 的概念表征方式优于 CCP。

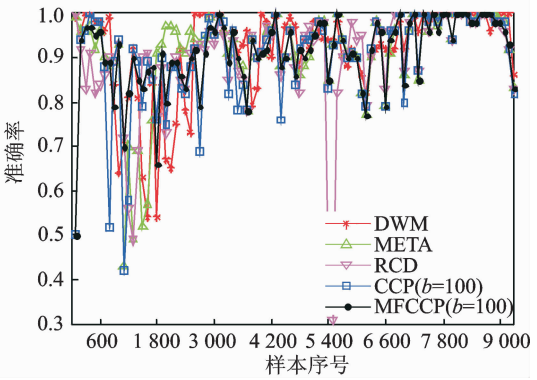


图 4　在 spam dataset 上的准确率

Fig. 4　Accuracy over spam dataset

3　结束语

本文提出的 MFCCP 算法抽取样本集的主要特征及其影响因子,主要用于表征该样本集所属的概念,以识别重复出现的概念,为每个概念维持一个分类器且及时更新,并依据 Hoeffding 不等式选择最合适的分类器对当前样本集实施分类。检测概念漂移需要判断整个特征集上类别的后验概率变化,而当特征维数较大时直接计算整个特征集上类别的后验概率变化通常不可行,因此本文并没有直接计算整个特征集上类别的后验概率变化,而是通过累加单个特征上类别的后验概率变化来趋近它。实验结果表明:主要特征对类别的影响的变化能够反映出整个特征集上类别的后验概率变化,本文提出的 MFCCP 算法非常适用于含重现概念漂移的数据流分类。目前,MFCCP 被设计成适用于样本特征值为二值型的数据,未来计划将其拓展使其适用于实型数据。

参考文献:

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

[2] Alexey T. The problem of concept drift: Definitions and related work [R]. Dublin: University of Dublin, Trinity College, Department of Computer Science, 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] Gama J, Zliobaite I, Bifet A, et al. A survey on concept drift adaptation [J]. ACM Computing Surveys, 2014, 46(4):1-35.

[5] Zliobaite I. Learning under concept drift: An overview [R]. Artificial Intelligence, Vilnius : Vilnius University, 2009.

[6] 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.
Wang Tao, Li Zhoujun, Yan Yuejin, et al. A survey of classification of data streams [J]. Journal of Computer Research and Development, 2007, 44(11): 1809-1815.

[7] 文益民, 强保华, 范志刚. 概念漂移数据流分类研究综述[J]. 智能系统学报, 2013, 8(2): 96-104.
Wen Yimin, Qiang Baohua, Fan Zhigang. A survey of the classification of data streams with concept drift [J]. CAAI Transactions on Intelligent Systems, 2013, 8(2): 96-104.

[8] Hoens T, Polikar R, Chawla N. Learning from streaming data with concept drift and imbalance: An overview [J]. Progress in Artificial Intelligence, 2012,1(1): 89-101.

[9] Dongre P, Malik L. A review on real time data stream classification and adapting to various concept drift scenarios [C]// Advance Computing Conference. Gurgaon, India: IEEE, 2014: 533-537.

[10] Katakis I, Tsoumakas G, Vlahavas I. Tracking recurring contexts using ensemble classifiers: An application to email filtering [J]. Knowledge and Information Systems, 2010, 22(3): 371-391.

[11] Read J, Bifet A, Pfahringer B, et al. Batch-incremental versus instance-incremental learning in dynamic and evolving data [J]. Advances in Intelligent Data Analysis, 2012, 7619: 313-323.

[12] 郭躬德,李南,陈黎飞. 一种基于混合模型的数据流概念漂移检测算法[J]. 计算机研究与发展, 2014, 51(4): 731-742.
Guo Gongde, Li Nan, Chen Lifei. Concept drift detection for data streams based on mixture model [J]. Journal of Computer Research and Development, 2014, 51(4): 731-742.

[13] Klinkenberg R . Learning drifting concepts: Example selection vs example weighting [J]. Intell Data Anal, 2004, 8(3):200-281.
[14] Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts[J]. Machine Learning, 1996, 23(1): 69-101.
[15] Sripirakas S, Russel P, Albert B, et al. Use of ensembles of Fourier spectra in capturing recurrent concepts in data streams [C]// International Joint Conference on Neural Networks . Killarney, Ireland:IEEE , 2015: 1-8.
[16] Mohammad J, Zahra A, Hamid B. Using a classifier pool in accuracy based tracking of recurring concepts in data stream classification[J]. Evolving Systems, 2013, 4(1):43-60.
[17] Ying Yang, Wu Xindong, Zhu Xingquan. Mining in anticipation for concept change: Proactive-reactive prediction in data streams [J]. Data Mining and Knowledge Discovery, 2006 , 13(3): 261-289.
[18] Ramamurthy S, Bhatnagar R. Tracking recurrent concept drift in streaming data using ensemble classifiers [C]// Proceedings of the Sixth International Conference on Machine Learning and Applications. Cincinnati, USA: IEEE, 2007: 404-409.
[19] Gama J, Kosina P. Tracking recurring concepts with meta-learners [J]. Progress in Artificial Intelligence, 2009, 5816: 423-434.
[20] Gonçalves P M, Barros R S. RCD: A recurring concept drift framework[J]. Pattern Recognition Letters, 2013, 34(9): 1018-1025.
[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. New York, USA:ACM, 2000:71-80.
[22] 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): 1078-1084
Zhu Qun, Zhang Yuhong, Hu Xuegang, et al. A double-window-based classification algorithm for concept drifting data streams[J]. Acta Automatica Sinica, 2011, 37(9): 1078-1084.
[23] Street W N, Kim Y. A streaming ensemble algorithm (SEA) for large-scale classification[C]// Proceedings of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2001: 377-382.
[24] Kolter J Z, Maloof M A. Dynamic weighted majority: An ensemble method for drifting concepts[J]. Journal of Machine Learning Research, 2007, 8: 2755-2790.
[25] Gama J, Medas P, Castillo G, et al. Learning with drift detection [C] // Proceedings of the Seventh Brazilian Symposium on Artificial Intelligence. Sao Luís, Brazil: Springer, 2004: 286-295.

作者简介:

冯超(1989-),男,硕士研究生,研究方向:数据挖掘,E-mail: henryfung01 @ 126.com。

文益民(1969-),男,教授,研究方向:数据挖掘、机器学习和社会计算。

汤凌冰(1975-),男,副教授,研究方向:数据挖掘和机器学习。