基于特征项分布的信息熵及特征动态加权概念漂移检测模型

孙 雪[1],李昆仑[2],韩 蕾[1],白晓亮[1]
(1.河北大学工商学院,河北保定 071002; 2.河北大学电子信息工程学院,河北保定 071000)

摘 要: 现有的概念漂移算法大多建立在数据流的分类模型上,忽略了特征空间与样本空间的分布特点,以及特征选择和加权的重要性.针对此问题提出了一种基于特征项分布的信息熵及特征动态加权算法,从概念漂移的动态演化性出发,根据样本和特征空间的拟合程度,运用特征信息熵理论对数据流中的概念漂移现象进行捕捉,以实现新旧概念的过渡.利用改进的隐含 Dirichlet 模型特征动态加权算法,以解决当前特征与历史特征的权重确定和无效特征的裁剪问题.在公开的语料库 CCERT 和 Trec06 上的测试实验证明了所提出算法的有效性.

关键词: 概念漂移; LDA 模型; 特征项分布; 信息熵
中图分类号: TP391.4; TP181 **文献标识码:** A **文章编号:** 0372-2112 (2015)07-1356-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.07.016

Construction of the Concept Drift Detection Model Based on the Information Entropy of Feature Distribution and Dynamic Weighting Algorithm

SUN Xue[1], LI Kun-lun[2], HAN Lei[1], BAI Xiao-liang[1]
(1. *Industrial and Commercial College, Hebei University, Baoding, Hebei* 071000, *China*;
2. *College of Electronic and Information Engineering, Hebei University, Baoding, Hebei* 071002, *China*)

Abstract: Most of the existing concept drift algorithm focuses on the classification model data streams, some of which overlook the distribution of the feature space and sample space, and the importance of feature selection and weighting. To solve this problem, we propose a dynamic information entropy and feature weighting algorithm based on the distribution of feature items from the dynamic evolution of the concept drift departure. To realize the concept transition, we capture the concept drifting of the data stream by the information entropy, according to the fitness degree between the sample and feature space. We improve the feature dynamic weighting latent dirichlet model, to overcome the problem of the current and historical feature weight assignment, as well as cropping the invalid features. Furthermore, the validity of the proposed algorithm was confirmed by the test in open corpus CCERT and Trec06.

Key words: concept drift; latent dirichlet allocation(LDA); feature distribution; information entropy

1 引言

近年来,随着信息技术的高速发展,数据流作为承载信息的重要媒介受到越来越广泛的关注,其与生俱有的实时性、随机性、多样性和开放性等特点,为数据的分析带来极大挑战[1].概念漂移(concept drift)是大规模数据流中存在的普遍现象,它是指随着时间的推移,数据流内部发生动态变化,使得原有训练数据与应用数据发生不匹配的现象.概念漂移问题的提出为数据流的优化决策提供了一条有效的途径,其相关研究在搜索引擎,入侵检测,网络安全等领域得到了较多的关注[2~4].如何准确快速捕捉数据流随时间的动态变化是研究概念漂移问题的核心内容,具体表现在:(1)在概念漂移的检测方面,如何提供快捷有效的机制来指导系统能够自适应的对概念漂移现象进行感知,监测和处理;(2)在新旧特征选择方面,如何提供有效的加权算法,达到新概念

收稿日期:2014-01-03;修回日期:2014-06-06;责任编辑:覃怀银
基金项目:国家自然基金(No.61073121);河北省自然基金(No.F2013201170);国家科技支撑计划(No.2013BAK07B04);河北省高等学校科学技术研究重点项目(No.ZD2014008);河北大学青年基金(No.2010Q17)

特征替换旧特征的目的;(3)在数据分析模型的构建方面,如何为漂移数据和原有数据之间的平稳过渡提供模型对接,使系统具备一定的感知能力和自适应能力.

早期的基于概念漂移的研究主要集中在单分类器的检测上,随着研究的不断深入,现有的算法已不再局限于单分类器设计,而更多的着眼于多分类器的集成学习[5].文献[6]提出了一种集成分类器模型,解决了不平衡数据流的概念漂移问题.文献[7]采用集成增量式学习概念漂移模型,根据当前和过去时刻分类器的错误率,动态更新样本权重.文献[8]利用随机决策树模型构建集成分类器,通过双层窗口机制周期性的检测滑动窗口中流数据的分布变化以适应数据漂移.随着对概念漂移问题的持续关注,一些新的方法不断出现[9].文献[10]提出了基于统计分布的模拟递推和集成算法,利用样本集的先验知识来选取最佳分类模型,实现了基于非监督学习的概念漂移检测.文献[11]采取循环监测的方式对感知数据流的突变有很好的效果.文献[12]采用朴素贝叶斯算法初始化训练样本权重,利用聚类结果计算节点阈值,根据决策树分类精度来重新设定集成分类器的权重.文献[13]选取支持向量机(Support Vector Machine,SVM)的估计作为时间窗口选择标准,通过时间指数衰减函数对样本进行加权.文献[14]提出了基于熵的概念漂移检测方法,通过计算新旧样本窗口中各个类别分布的熵来评测训练集之间样本分布的差异.

目前针对概念漂移问题的研究主要集中在两方面[5~13],一是利用更新样本权重的方法来实现漂移前后训练集的更新;二是算法多数建立在数据流的分类模型上,通过分类精度的变化来感知漂移现象.随着时间的推移样本空间发生改变,原有特征空间无法表征当前数据模型,在原有模型上单纯依靠分类决策难以提高算法的准确性.针对上述问题,本文从概念漂移的动态演化性出发,利用特征在样本空间和特征空间的分布特点对数据流中的概念漂移现象进行捕捉,提出了基于特征项分布的信息熵概念漂移检测模型(Information Entropy of Feature Distribution,IEFD),并利用改进的 LDA 特征动态加权算法(Feature Dynamic Weighting assignment LDA model,FDW-LDA)实现新旧概念的过渡,以解决当前特征与历史特征的权重衡量和无效特征的裁剪问题.

2　基于特征项分布的信息熵概念漂移检测模型

2.1　问题描述

随着输入数据的不断更新,当前样本空间 $f(S)$ 与特征空间 $f(T)$ 的信息匹配度降低,原有特征选择模型将不再适合当前样本的分布特点,使得分类精度降低,如图 1 所示.图中左侧椭圆表示历史数据经过特征选择筛选出的特征,右侧椭圆表示当前数据流包含的所有特征,图中交叉部分为有效特征集合.当概念漂移现象发生后,有效特征数目减少,特征空间将不能很好的拟合现有数据流的构成.因此利用样本分布特点,根据特征中隐藏的信息,及时寻找特征空间变化规律是解决概念漂移问题的有效途径.在进行算法描述之前,首先给出如下定义.

定义 1　设当前样本空间为 $f(S)$,数据集合 Ω 中每条数据都由若干个特征组成,假设当前数据流包含有效特征项的个数为 d,则样本空间特征个数为 $a+d$.

定义 2　特征空间为 $f(T)$,任意一组数据 x 的特征向量表示为 $f_x(T)=(w_1,w_2,\cdots,w_d,0,\cdots,0)$,其中 0 为不包含的特征项个数 b,则特征空间含有的特征项个数为 $b+d$.

定义 3　特征空间与样本空间的拟合度为 H,用样本空间信息熵 $H_{f(S)}$ 和特征空间信息熵 $H_{f(T)}$ 来评估特征项的流动性能,反馈两个空间的相容性.

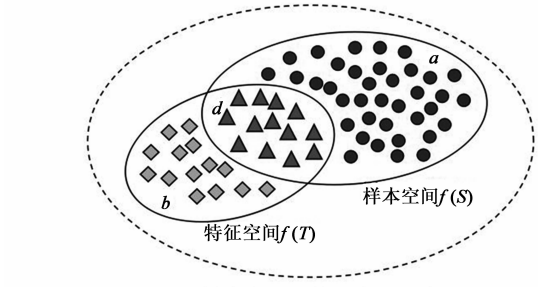


图1　样本空间与特征空间示意图

2.2　基于特征项分布的信息熵概念漂移检测算法

信息熵[15]是以数学的方法来度量去除冗余信息干扰后,事件所提供的平均信息量,为抽象信息量化提供了参考依据.本文提出的基于特征项分布的概念漂移检测算法,主要是从样本空间和特征空间来寻找特征信息熵的变化规律,根据特征项的匹配度来预测样本未来的发展趋势,感知概念漂移现象的发生.下式分别给出了样本空间信息熵 $H_{f(S)}$ 和特征空间信息熵 $H_{f(T)}$ 的描述方法,$i=1(j=1)$ 表示样本(特征)空间包含的特征项分布,$i=2(j=2)$ 表示不包含的特征项分布.

$$H_{f(S)}=-\sum_{i=1}^{2}p(c_i)\log_2 p(c_i)$$
$$=-\left(\frac{d}{a+d}\right)\log_2\left(\frac{d}{a+d}\right)-\left(\frac{a}{a+d}\right)\log_2\left(\frac{a}{a+d}\right)\quad(1)$$

$$H_{f(T)}=-\sum_{j=1}^{2}p(c_j)\log_2 p(c_j)$$
$$=-\left(\frac{d}{b+d}\right)\log_2\left(\frac{d}{b+d}\right)-\left(\frac{b}{b+d}\right)\log_2\left(\frac{b}{b+d}\right)\quad(2)$$

特征项在样本空间和特征空间的分布特点,决定

了样本与特征空间的拟合程度,为将其量化,需要针对 a、b、d 三者的数量关系分如下三种情况讨论,对信息熵 H 进行整合.

(1)当 $a>d\wedge b>d$ 且 $d\neq 0$ 或 $a<d\wedge b<d$ 时

$$H=\delta+\sigma(H_{f(T)}+H_{f(S)})=\delta-\sigma\Big[\sum_{i=1}^{2}p(c_i)\log_2 p(c_i)+\sum_{j=1}^{2}p(c_j)\log_2 p(c_j)\Big] \tag{3}$$

(2)当 $a<d<b$,且 $a\neq 0$ 时

$$H=\delta+\sigma H_{f(T)}+H_{f(S)}=\delta-\sigma\sum_{i=1}^{2}p(c_i)\log_2 p(c_i)-\sum_{j=1}^{2}p(c_j)\log_2 p(c_j) \tag{4}$$

(3)当 $b<d<a$ 时

$$H=\delta+H_{f(T)}+\sigma H_{f(S)}=\delta-\sum_{i=1}^{2}p(c_i)\log_2 p(c_i)-\sigma\sum_{j=1}^{2}p(c_j)\log_2 p(c_j) \tag{5}$$

补偿因子 δ,当样本中包含特征项的数量 d 较大,超过设定阈值时需要对熵值进行补偿.

互补系数 σ,均衡样本空间与特征空间的差异值,$\sigma\in[-1,1]$.

如果特征空间可以精确地刻画出输入样本集的区分度信息,则说明两个空间信息匹配,反之则说明发生了概念漂移,两个空间存在描述偏差,导致模型分类精度下降,如算法 1.

算法 1 基于特征项分布的信息熵概念漂移检测算法 IEFD

Input:数据流 Ω 包含 m 时刻前所对应的数据序列:
$C_\Omega=\{C_1,C_2,\cdots,C_{m-1},C_m\}$,特征项$\{w_1,w_2,\cdots,w_{b+d}\}$;

Output:测试集合的信息熵权重 H;

Begin:

Step1 初始化,数据预处理;

Step2 计算 m 时刻数据 C_m所包含样本空间$f(S)$特征的个数 $a+d$;

Step3 将 $f(S)$特征与特征项$\{w_1,w_2,\cdots,w_{b+d}\}$对比,得出有效特征个数 d;

Step4 生成向量空间模型,将样本空间 $f(S)$转换成特征空间 $f(T)$;

Step5 利用式(1)、(2),计算样本空间信息熵 $H_{f(S)}$和特征空间信息熵 $H_{f(T)}$;

Step6 通过 a、b、d 三者的数量关系,利用式(3)~(5),得出测试集合的信息熵权重 H;

End

3 LDA 特征动态加权概念漂移算法

3.1 基于 LDA 的特征动态加权算法

LDA(Latent Dirichlet Allocation)[16,17]是近几年新兴的一种数据处理算法,它采用概率主题结构对数据进行建模,生成一个三层贝叶斯网络,各层结构依次为数据集层$\{D_1,D_2,\cdots,D_n\}$,主题层$\{T_1,T_2,\cdots,T_k\}$,和特征层$\{w_1,w_2,\cdots,w_n\}$,特征词 w 的权重由主题-词的条件概率分布 $p(w_i\mid T_1,T_2,\cdots,T_k)$表示,计算公式如下:

$$W_{\mathrm{LDA}}=p(w_i\mid T_1,T_2,\cdots,T_k)=\frac{n_{\vec{i},j}^{(w_i)}+\beta}{n_{\vec{i},j}^{(-)}+V\beta} \tag{6}$$

符号 V 表示样本空间特征项的总数;$n_{\vec{i},j}^{(w_i)}$表示在主题 T_j下生成所有特征的个数;β 为 Dirichlet 分布参数,可通过 Gibbs 抽样得到,$n_{\vec{i},j}^{(-)}$表示在所有主题下生成特征项的总数.式(6)中分子与分母项差值较大,尤其针对大数据集这种现象越发明显,对特征项的重要程度难以做出正确的判断,而且式中仅定义了特征 w_i 属于某一特定主题的概率,缺乏考虑不同主题生成相同特征的情况,因此给出一种改进的特征词权重公式如下:

$$W=\lambda-\sum_{j=1}^{K}\lg\left(\frac{n_{\vec{i},j}^{(w_i)}+\beta}{n_{\vec{i},j}^{(-)}+V\beta}\right) \tag{7}$$

其中 λ 为调节参数,可根据具体的数据集进行调整.

3.2 特征动态加权概念漂移算法

图 2 给出了随着数据种类的变化,新特征注入样本空间,产生概念漂移现象的示意图.图左侧表示历史数据特征空间,右侧表示产生概念漂移后的当前特征空间,图中圆点表示特征集合,连线表示各个特征间的互相联系.如图所示当数据流发生改变后,特征空间将出现无效特征集和新特征集,更新现有特征空间可以有效提高算法准确度.

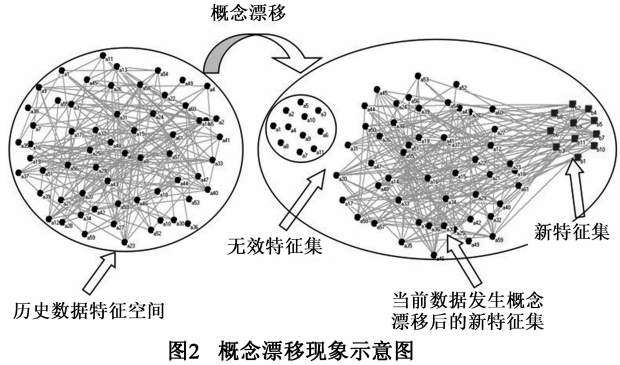


图2 概念漂移现象示意图

在特征动态加权概念漂移算法中,新旧特征的取舍是影响算法准确性的关键因素,而历史和当前数据的平均特征权重值是对特征重要性的客观反馈.因此两者之间的比例关系可以作为新旧特征权重的指导项.

历史数据集用下标 h 表示,新增数据用 a 表示.r 代表数据中包含特征项的个数,$\overline{W_{a(\mathrm{LDA})}}$和$\overline{W_{h(\mathrm{LDA})}}$分别代表新增特征和历史特征的 LDA 权重平均值,根据式(6)得出其计算式,如下:

$$\overline{W_{a(\mathrm{LDA})}}=\frac{1}{r_a}\sum_{i=1}^{r_a}\left(\frac{n_{a(\vec{i},j)}^{(w_i)}+\beta_a}{n_{a(\vec{i},j)}^{(-)}+V_a\beta_a}\right) \tag{8}$$

$$\overline{W_{h(\mathrm{LDA})}} = \frac{1}{r_h}\sum_{i=1}^{r_h}\left(\frac{n_{h(i,j)}^{(w_i)} + \beta_h}{n_{h(i,j)}^{(-)} + V_h\beta_h}\right) \tag{9}$$

根据式(7),新增特征权重 W_a 的计算公式如下.在式中加入了新旧特征权重指导项 $\overline{W_{a(\mathrm{LDA})}}/\overline{W_{h(\mathrm{LDA})}}$,该项对概念漂移的趋势起引导作用,通过新旧特征权重值对比,可实现去除冗余特征,扩充新特征集的目的,如算法 2.

$$W_a = \lambda_a - \overline{W_{a(\mathrm{LDA})}}/\overline{W_{h(\mathrm{LDA})}}\sum_{j=1}^{K}\lg\left(\frac{n_{a(i,j)}^{(w_i)} + \beta_a}{n_{a(i,j)}^{(-)} + V_a\beta_a}\right)$$

$$= \lambda_a - \frac{r_h\sum_{i=1}^{r_a}\left(\frac{n_{a(i,j)}^{(w_i)} + \beta_a}{n_{a(i,j)}^{(-)} + V_a\beta_a}\right)\sum_{j=1}^{K}\lg\left(\frac{n_{a(i,j)}^{(w_i)} + \beta_a}{n_{a(i,j)}^{(-)} + V_a\beta_a}\right)}{r_a\sum_{i=1}^{r_h}\left(\frac{n_{h(i,j)}^{(w_i)} + \beta_h}{n_{h(i,j)}^{(-)} + V_h\beta_h}\right)} \tag{10}$$

算法 2　基于 LDA 模型的特征动态加权算法 FDW-LDA

Input: m 时刻发生概念漂移后的数据序列 C_m,主题数 K 以及每个主题所包含特征项的个数 r;

Output:当前数据集特征空间矩阵 $\boldsymbol{B}$

Begin:

Step1　在数据集 C_m 上重构 LDA 主题模型;

Step2　计算 $m-1$ 时刻前 $\{C_1, C_2, \cdots, C_{m-1}\}$ 序列的历史特征权重和 C_m 新增特征 LDA 权重 $\{W_{a1(\mathrm{LDA})}, W_{a2(\mathrm{LDA})}, \cdots, W_{an(\mathrm{LDA})}\}$;

Step3　根据式(8)、(9),计算新增特征和历史特征的 LDA 权重平均值 $\overline{W_{a(\mathrm{LDA})}}, \overline{W_{h(\mathrm{LDA})}}$;

Step4　利用式(10),计算新增特征权重 W_a;

Step5　依照权重值大小关系将 W_a 和 W_h 排序,按照经验比例扩充新增特征;

Step6　删除无效特征和冗余特征,更新特征空间矩阵 $\boldsymbol{B}$;

End

4　实验

4.1　实验数据及评价指标

垃圾邮件具有形式多样,实时动态性强等特点,非常适合于验证概念漂移算法的有效性,因此本文实验数据选取两个公开的邮件集,中国教育和科研计算机网紧急响应组(China education and research network Computer Emergency Response Term, CCERT)提供的数据集和国际文本检索会议提供的 Trec06 语料库.邮件的预处理包括分词,词性选择,过滤常用词,删除高频低频词.

4.2　基于特征项分布的信息熵概念漂移检测算法

实验中,特征选择采用开方拟合检验 CHI 算法提取 450 维特征向量.图 3 中给出了基于特征项分布的特征熵概念漂移 IEFD 算法在 Trec06c 语料库上的系统精确率(accuracy),测试数据集包括 5980 篇正常邮件和 11910 篇垃圾邮件,按照测试数据与训练数据集的接近程度将其分成 6 组数据,用来模拟概念漂移现象的发生.实验采用三种分类器,分别是多项式贝叶斯(Multinomial Naive Bayes, Mul-NB),K 近邻(K-Nearest Neighbor, K-NN)和 Adaboost M1.第 1 组数据完全选自新类型邮件,共包括 140 封邮件,通过实验计算得出该组数据中包含特征项的个数与文本中所有词的个数比例小于等于 0.05,且文本中包含特征项的个数与特征项总数之比小于 0.03,该组数据的平均信息熵权重为 0.21,其在三种分类器下的分类精度也处于最低值.从第 1 组数据到第 6 组数据,测试数据与训练数据集的接近程度逐渐升高,信息熵权重与精确率也在逐步提高.

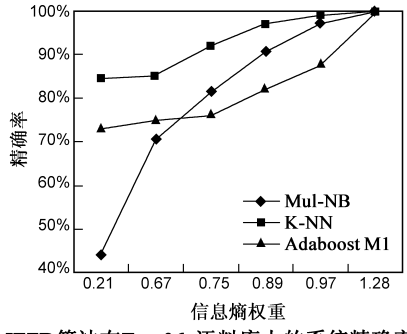


图3　IEFD算法在Trec06c语料库上的系统精确率对比图

图 4 中测试数据集包括 2271 篇正常邮件和 6255 篇垃圾邮件,为细化实验结果,将测试数据分成 8 组,分组方法与图 3 相同.为保证实验结果的泛化性将上组实验中的 Adaboost M1 更换为 J48 决策树.通过曲线对比可知,测试数据与训练数据集的匹配度越高,信息熵权重值越高,分类的精确率也越高,发生概念漂移的几率越低.通过在不同数据集上的两个实验对比可以看出,信息熵权重,测试文本与训练集的相似度,系统的精确率这三者存在同向关系,信息熵权重可以作为检测概念漂移现象是否发生的量化指标.

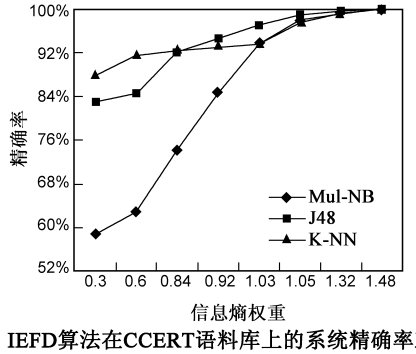


图4　IEFD算法在CCERT语料库上的系统精确率对比图

4.3　LDA 特征动态加权概念漂移算法

从数据集中随机抽取 45% 作为训练数据,其余 55% 作为测试数据,垃圾邮件和正常邮件的比例为2:1,

利用 LDA 特征加权函数选取 450 维特征词.实验所用分类器为 Mul-NB,K-NN 和标准支持向量分类(C-Support Vector Classification,C-SVC).

表 1、2 记录了各种特征加权算法在 Trec06c 和 CCERT 数据集上的系统精确率对比结果.实验将 FDW-LDA 与文本分类中常用的特征选择和加权算法做了对比,系统精确率是衡量邮件过滤系统性能的一个重要指标,从表中可以看出,与其它 7 种特征加权算法相比,基于 LDA 特征动态加权算法具有优势,在不同数据集和分类器上分类差异较小,性能稳定.

表 1 在 Trec06c 数据集上各种特征加权算法的系统精确率对比结果

特征权重选择函数	Mul-NB	C-SVC	K-NN
开方拟合检验 Chi(Chi-Square)	90.35%	93.73%	95.66%
特征熵 TE(Term Entropy)	92.98%	80.44%	95.97%
互信息 MI(Mutual Information)	94.55%	91.61%	96.16%
词频 TF(Term Frequency)	93.76%	94.91%	95.69%
词频-文档频度 TF-IDF (-Inverse Document Frequency)	92.92%	95.03%	95.88%
权重归一化 LTC(Log TF-IDF)	92.02%	93.96%	91.68%
LDA 特征动态加权算法 FDW-LDA	95.07%	96.59%	96.76%

表 2 在 CCERT 数据集上各种特征加权算法的系统精确率对比结果

特征权重选择函数	Mul-NB	C-SVC	K-NN
布尔向量化	93.75%	97.08%	89.23%
开方拟合检验 Chi	83.03%	93.27%	94.64%
词频 TF	92.65%	97.36%	95.84%
词频-文档频度 TF-IDF	92.89%	95.83%	95.83%
权重归一化 LTC	96.42%	96.78%	94.28%
LDA 特征动态加权算法 FDW-LDA	96.95%	97.78%	97.43%

图 5、6 给出了特征加权算法在 CCERT 和 Trec06c 上的垃圾邮件准确率(spam precision)对比结果,在 Trec06c 数据集上,FDW-LDA 算法的实验效果相对较为理想,没有受到不同分类器的影响,而 TE 和 LTC 算法受分类器影响较大,Chi、TF 和 TF-IDF 算法性能相对比较稳定,实验结果接近.

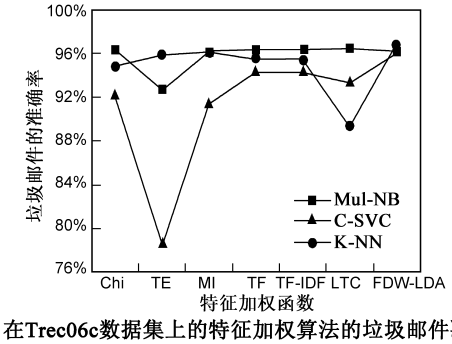


图5 在Trec06c数据集上的特征加权算法的垃圾邮件准确率对比图

如图 6 所示,在 Multinomial-Naive Bayes 分类器上,各种加权算法实验结果比较平稳,准确率在 97.6% ~ 98.88%之间波动,TF 最高,FDW-LDA 最低,与最高值仅差 1.28%,差别较小.在 K-NN 分类器上,FDW-LDA 的准确率最高,比最低值高出 8%,从三条曲线的变化趋势和拐点位置可以看出,FDW-LDA 算法性能平稳没有较大波动,而其它特征加权算法在不同分类器上的性能表现变化较大,如 Chi 算法在 C-SVC 分类器表现最好,而在 K-NN 分类器上算法准确率相对较低.上述实验说明 FDW-LDA 算法具有良好的泛化性能.

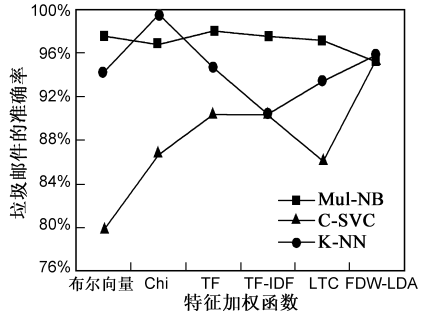


图6 在CCERT数据集上的特征加权算法的垃圾邮件准确率对比图

5 结论

概念漂移问题作为数据流中存在的普遍现象,成为数据挖掘领域的研究热点,但由于数据流本身具有动态性,不确定性,不可预知性,多变性等特点,致使针对概念漂移现象的检测成为难点.本文研究了基于特征项分布的信息熵及特征动态加权概念漂移检测模型,提出了两个算法,其创新之处主要包括:①提出的基于特征项分布的信息熵概念漂移算法,从样本空间和特征空间寻找特征信息熵的变化规律,根据特征项的匹配度来预测样本未来的发展趋势,感知概念漂移现象的发生;②提出的基于 LDA 模型的特征动态加权算法,采用重构主题模型的方式搜寻新特征,达到扩充原有特征集和去除冗余特征的目的,提高了系统的性能,具有普遍意义.下一步考虑将本文设计的概念漂移模型运用到其它数据流中,结合不同数据流的结构特点,提高算法的泛化性.

参考文献

[1] WIDMER G,KUBAT M. Learning in the presence of concept drift and hidden contexts[J]. Machine Learning,1996,(23):69 - 101.

[2] HOENS T R,POLIKAR R,et al. Learning from streaming data with concept drift and imbalance: an overview[J]. Progress in

Artificial Intelligence,2012,1(1):89 – 101.

[3] 文益民,强保华,等.概念漂移数据流分类研究综述[J].智能系统学报,2013,8(2):95 – 104.
WEN Yi-min,QIANG Bao-hua,et al. A survey of the classification of data streams with concept drift [J]. CAAI Transactions on Intelligent Systems,2013,8(2):95 – 104.(in Chinese)

[4] 柴玉梅,张卓,等.基于频繁概念直乘分布的全局闭频繁项集挖掘算法[J].计算机学报,2012,35(5):990 – 1000.
CHAI Yu-mei, ZHANG Zhuo, et al. An algorithm for mining global closed frequent itemsets based on distributed frequent concept direct product [J]. Chinese Journal of Computers, 2012,35(5):990 – 1000.(in Chinese)

[5] 孙岳,毛国君,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.
SUN Yue,MAO Guo-jun,et al. Mining concept drifts from data streams based on multi-classifiers[J]. Acta Automatica Sinica,2008,34(1):93 – 97.(in Chinese)

[6] 欧阳震诤,罗建书,等.一种不平衡数据流集成分类模型[J].电子学报,2010,38(1):184 – 189.
OUYANG Zhen-zheng,LUO Jian-shu,et al. An ensemble classifier frame work for mining imbalanced data streams[J]. Acta Electronica Sinica,2010,38(1):184 – 189.(in Chinese)

[7] ELWELL R,POLIKAR R. Incremental learning of concept drift in non-stationary environments [J]. IEEE Transactions on Neural Networks,2011,22(10):1517 – 1531.

[8] 朱群,张玉红,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,2011,9(37):1077 – 1084.
ZHU Qun, ZHANG Yu-hong, et al. A double-window-based classification algorithm for concept drifting data streams[J]. Acta Automatica Sinica, 2011, 9(37): 1077 – 1084.(in Chinese)

[9] 徐文华,覃征,常扬.基于半监督学习的数据流集成分类算法[J].模式识别与人工智能,2012,25(2):292 – 299.
XU Wen-hua, QIN Zheng, CHANG Yang. Semi-supervised learning based ensemble classifier for stream data[J]. Pattern Recognition and Artificial Intelligence,2012,25(2):292 – 299.(in Chinese)

[10] PIOTR S,MICHAL W. Concept drift detection and model selection with simulated recurrence and ensembles of statistical detectors[J]. Journal of Universal Computer Science,2013,19(4):462 – 483.

[11] PAULO M G, ROBERTO S M. RCD: A recurring concept drift framework[J]. Pattern Recognition Letters, 2013, 34(9):1018 – 1025.

[12] DEWAN M F,LI Z,et al. An adaptive ensemble classifier for mining concept drifting data streams[J]. Expert Systems with Applications,2013,40(15):5895 – 5906.

[13] KLINKENBERG R. Learning drifting concepts: example selection vs. example weighting [J]. Intelligent Data Analysis, 2004,8(3):281 – 300.

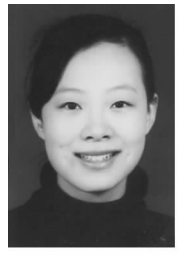
[14] PETER V, ABRANHAM B. Entropy-based concept drift detection[A]. Han J W. Proceedings of the 6th International Conference on Data Mining[C]. Houston: IEEE Computer Society,2006.1113 – 1118.

[15] 于剑,石洪波,等.关于极大熵聚类算法的收敛性定理的反例[J].中国科学 E 辑:技术科学,2003,33(6):531 – 536.

[16] BLEI D M, NG A Y, et al. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, (3): 993 – 1022.

[17] 石晶,范猛,等.基于 LDA 模型的主题分析[J].自动化学报,2009,35(12):1586 – 1592.
SHI Jing, FAN Meng, et al. Topic analysis based on LDA model[J]. Acta Automatica Sinica, 2009, 35(12): 1586 – 1592.(in Chinese)

作者简介



孙　雪　女,1981 年 1 月出生于天津市.现为河北大学工商学院老师.从事模式识别与人工智能,机器学习与数据挖掘,信息安全方面的研究.
E-mail:sunxue@hbu.edu.cn

李昆仑　男,1962 年 7 月出生于河北保定市.教授,硕士生导师.主要研究方向为模式识别与人工智能,机器学习与数据挖掘,信息安全,生物信息技术等.
E-mail:likunlun@hbu.edu.cn