基于贝叶斯方法和变化表的恐怖行为预测算法

薛安荣 毛文渊 王孟頔 陈泉浈
(江苏大学计算机科学与通信工程学院 镇江 212013)

摘 要 传统的恐怖行为预测算法没有考虑到组织会改变其行为策略,而CAPE算法根据组织背景的改变预测行为变化,但其只能根据变化表中存在的背景变化预测行为。为了能根据任意背景变化预测恐怖行为,针对恐怖数据高维小样本的特点,提出了一种利用贝叶斯方法在改进的变化表上预测组织行为的算法。利用贝叶斯方法可快速有效地解决高维小样本分类问题的特性,在改进的变化表上实现对组织行为的预测,从而提高了预测精度和计算效率。此外,考虑到背景的变化会在时间序列上对组织行为产生持续的影响,因此在不同时间滞差下,利用加权的贝叶斯方法预测组织行为。MAROB数据集上多个组织数据的实验结果也表明,所提算法在准确率及时间复杂度上优于CAPE算法。

关键词 恐怖预测,贝叶斯方法,变化表,加权贝叶斯
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.023

Terrorism Prediction Based on Bayes Method and Change Table

XUE An-rong MAO Wen-yuan WANG Meng-di CHEN Quan-zhen
(School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract Traditional terrorism behavior prediction algorithms do not consider how the group will change its behaviors. CAPE predicts changes of behaviors according to context variation of organizations, but it only predicts the changes of behavior based on changes of the context, which is existed in its change table. Considering the characteristics of the high dimensions and small samples of terrorism data, this paper proposed a terrorism prediction algorithm based on improved change table using Bayes method, to predict organizational behavior according to any behavior changes. It predicts organization behaviors on the change table due to the fact that Bayes method classifies high dimensions and small sample in a fast and efficient way. Thus, it improves prediction precision and computing efficiency. In addition, considering the continuing effect of the change of the group's context on its behavior, the weighted Bayes method with different time lags is used to predict the behavior of the organization. Experiments on multiple organization data of MAROB show that, the proposed algorithm is better than CAPE algorithm on accuracy and time complexity.

Keywords Terrorism prediction, Bayes method, Change table, Weighted Bayes

恐怖袭击事件发生的原因包含政治、经济和宗教信仰等方面的因素,各种因素交织在一起,使恐怖行为预测显得错综复杂。现有的恐怖预测模型主要是根据组织以往背景与行为之间的联系来预测其未来的表现,并未考虑组织背景属性变化引起的行为属性的改变[1-4]。以文化模型(Cultural Modeling)[5]为基础的算法就是根据组织背景与行为之间的关系,构建组织的行为预测模型,如SOMA[6,7]、CONVEX[8]等算法,而组织一般具有反侦查能力,其恐怖活动发生的时间、地点及行为强度等属性会因此改变。现有模型大多数没有考虑组织的这些改变以及由此引起的行为变化。只有CAPE(Change Analysis Predictive Engine)模型[9]考虑了组织行为持续改变并动态变化的情况,其基本思想是组织背景的改变可能引起其行为的变化,通过学习组织改变行为的条件,构建背景与行为之间的变化规则。预测时,CAPE利用这些规则预测组织未来在什么样的背景变化下会改变它们的行为。然而该模型无法根据变化表之外的背景变化预测行为变化,必须结合SitCAST(Situation Forecast)和CONVEX(Context Vectors)方法来预测组织行为。但SitCAST+CONVEX方法预测准确率较低,且SitCAST算法是一个指数级的过程,CAPE的时间复杂度上限急剧增加。而在最坏的情况下,CAPE算法可能会退化为SitCAST+CONVEX算法,使算法的预测准确率较低而时间复杂度较高。此外,CAPE模型使用变化表和SitCAST+CONVEX算法组合预测恐怖行为,使算法的流程变得复杂。

为了能够在任意的背景变化下预测组织行为,统一预测过程,并且针对数据集高维小样本特点,本文提出一种基于贝叶斯方法和变化表的恐怖行为预测算法。改进的变化表存储了背景与行为已有的变化情况,使贝叶斯方法能方便地从中收集相关的变化信息。对于新的背景变化,贝叶斯方法根据输入背景变化向量中各个元素对不同分类结果的影响程度进

到稿日期:2015-11-15 返修日期:2016-03-05 本文受国家自然科学基金(61300228)资助。
薛安荣(1964-),男,博士,教授,CCF高级会员,主要研究方向为数据挖掘、机器学习,E-mail:xuear@mail.ujs.edu.cn;**毛文渊**(1989-),男,硕士生,主要研究方向为数据挖掘;**王孟頔**(1992-),女,硕士生,主要研究方向为数据挖掘;**陈泉浈**(1990-),男,硕士生,主要研究方向为数据挖掘。

行综合判断，从而能够在任意的背景变化下实现预测行为的目的。除此之外，对于高维小样本数据，贝叶斯方法能够快速有效地预测。最后，考虑到背景的改变对行为的影响不是瞬时的，而是会在一定的时间周期之内对行为产生持续影响，因此，算法考虑在不同时间滞差下，利用加权贝叶斯综合预测组织行为。

1 变化表

1.1 MAROB数据集

MAROB(少数极端组织表现)[10]数据集由美国国际发展和冲突管理中心(CIDCM)和美国国土安全部的MAR项目收集。该数据集提供了全球大多数恐怖组织的社会特征，其中包括组织的意识形态、宗教信仰、政治主张、经济情况等背景属性和武装袭击、自杀袭击等行为属性。若令 g 表示恐怖组织，Tg 表示数据集中组织的数据，则 $Tg=(x_1,x_2,\cdots,x_m)^{\mathrm{T}}$，$x_i=(c_1,c_2,\cdots,c_n,a_1,a_2,\cdots,a_k)$，其中 c_i 属于背景属性集 $CS(g)$，a_j 属于行为属性集 $AS(g)$。

MAROB数据集的主要特点如下：

1)高维小样本数据。MAROB数据集中记录了恐怖组织109种背景属性，而单个组织的可用信息最多只有25条，具有高维小样本的特点。

2)有用信息较少。MAROB数据集中各个组织的背景属性集中值不变的属性有一半左右，变化频度低于2次的背景属性大约占60%～70%。因此，背景属性集中有用信息较少。

香农信息熵理论[11]表明，越是有序的序列，其中蕴含的信息量越少。鉴于组织属性集中存在大量具有相同值的属性，在数据预处理阶段，使用简单高效的基于最大相关的特征选择方法[12]选择背景属性子集。

1.2 改进的变化表

为了更直观地表示组织行为变化与背景变化之间的联系，需要在背景数据集上构造出变化表[8]。变化表是一种记录组织在时间序列上行为及其背景变化关系的数据存储结构，表示背景的改变对行为变化的影响。原始的变化表被用来从中提取组织改变行为的规则，根据满足规则要求的背景变化来预测行为变化。它只能利用过去已有的背景变化，而不能使用任意的背景变化预测行为。

定义1(变化) 在时间序列上，属性前一时刻与当前时刻形成的值对，称为该属性在当前时刻的变化。

定义2(变化表) 令 Tg 表示数据集中组织 g 的表现数据，其中 $CS(g)$ 属于背景属性集，$AS(g)$ 属于行为属性集。变化表 $CT(g,A_j)$ 中时间段 i 的记录由背景属性从时间段 $i-1$ 到时间段 i 的变化以及行为属性从时间段 i 到时间段 $i+1$ 的变化产生。

如表1所列，包含背景和行为 A_j 的原始子数据经过处理后，生成如表2所列的行为 A_j 的变化表 $CT(g,A_j)$。

表1 原始数据

Time	C_1	C_2	…	C_n	A_j
1	$C_1[1]$	$C_2[1]$	…	$C_n[1]$	$A_j[1]$
2	$C_1[2]$	$C_2[2]$	…	$C_n[2]$	$A_j[2]$
3	$C_1[3]$	$C_2[3]$	…	$C_n[3]$	$A_j[3]$
…	…	…	…	…	…
m	$C_1[m]$	$C_2[m]$	…	$C_n[m]$	$A_j[m]$

表2 行为 A_j 的变化表 $CT(g,A_j)$

C_1	C_2	…	C_n	PA_j	LA_j
$(C_1[1],C_1[2])$	$(C_2[1],C_2[2])$	…	$(C_n[1],C_n[2])$	$A_j[2]$	$A_j[3]$
$(C_1[2],C_1[3])$	$(C_2[2],C_2[3])$	…	$(C_n[2],C_n[3])$	$A_j[3]$	$A_j[4]$
…	…	…	…	…	…
$(C_1[m-2],C_1[m-1])$	$(C_2[m-2],C_2[m-1])$	…	$(C_n[m-2],C_n[m-1])$	$A_j[m-1]$	$A_j[m]$

当预测行为时，将当前时刻的背景属性的变化及行为作为输入向量，预测下一个阶段的行为属性值。例如，在未来有输入向量$((C_1[1],C_1[2]),(C_2[1],C_2[2]),\cdots,(C_n[1],C_n[2]),A_j[2])$，那么预测下一时刻组织的行为将可能变为 $A_j[3]$。

从变化表中可以看到，背景属性从时间段 $i-1$ 到 i 的变化，只能用来学习行为属性从 i 到 $i+1$ 的变化。即以上的变化表只表明了背景变化与一个时间周期之后行为变化的关系。然而，背景的改变可能会在一段时间之后依然对行为的变化产生影响。因此，必须考虑背景变化与行为变化之间的影响时间滞差 h。

定义3(h-变化表) 令 Tg 表示数据集中组织 g 的表现数据，其中 $CS(g)$ 属于背景属性集，$AS(g)$ 属于行为属性集。对于整数 $h>1$，h-变化表 $CT^h(g,A_j)$ 中时间段 i 的记录由背景属性从时间段 $i-h$ 到 $i-h+1$ 的变化以及行为属性从 i 到 $i+1$ 的变化产生。

如表1所列，包含背景和行为 A_j 的原始数据经过处理后，生成如表3所列的行为 A_j 的 h-变化表 $CT^h(g,A_j)$。

表3 行为 A_j 的 h-变化表 $CT^h(g,A_j)$

C_1	C_2	…	C_n	PA_j	LA_j
$(C_1[1],C_1[2])$	$(C_2[1],C_2[2])$	…	$(C_n[1],C_n[2])$	$A_j[h+1]$	$A_j[h+2]$
$(C_1[2],C_1[3])$	$(C_2[2],C_2[3])$	…	$(C_n[2],C_n[3])$	$A_j[h+2]$	$A_j[h+3]$
…	…	…	…	…	…
$(C_1[m-h-1],C_1[m-h])$	$(C_2[m-h-1],C_2[m-h])$	…	$(C_n[m-h-1],C_n[m-h])$	$A_j[m-1]$	$A_j[m]$

h-变化表的构造定义中只添加了一个参数 h，代表了背景属性变化与行为属性变化关联的时间滞差。

2 基于贝叶斯方法的恐怖行为预测

传统的机器学习方法在高维小样本数据上往往难以达到令人满意的效果[13]，而贝叶斯方法在高维小样本数据上能获得较快的分类速度及准确度。对于训练样本中没有的条件元组，贝叶斯方法也能根据各个属性值对不同分类结果的影响程度进行分类，从而能预测属性集中属性取值的所有组合。

2.1 贝叶斯定理

贝叶斯定理是一种把类的先验知识和从数据集中收集的类的后验知识相结合的统计学方法[14]。

$$P(C|X)=\frac{P(X|C)P(C)}{P(X)} \tag{1}$$

其中，$P(C|X)$ 是后验概率，即在条件 X 下 C 发生的概率；$P(C)$ 是 C 的先验概率。很显然，贝叶斯定理提供了一种由 $P(X)$，$P(C)$ 和 $P(X|C)$ 来计算后验概率 $P(C|X)$ 的方法。

分类时，假定 $C_1,C_2,\cdots,C_m$ 为 m 个分类的类标号，X 为输入元组向量，那么 $P(C_i|X)$ 表示在输入元组为 X 的条件下，分类结果为 C_i 的概率。贝叶斯分类的结果就是使 $P(C_i|X)$ 值最大的类 C_i，而对于相同的元组 X，它的概率 $P(X)$ 总是相同的，因此只需选择使 $P(X|C_i)P(C_i)$ 概率值最大的类 C_i，即贝叶斯分类法预测 X 属于类 C_i，当且仅当

$$P(C_i|X)P(C_i)>P(C_j|X)P(C_j),1\leqslant j\leqslant m,j\neq i \quad (2)$$

其中，$P(C_i)$根据训练数据中类 C_i 的概率获得。在类属性相互独立的假定下，$P(X|C_i)$可以通过式(3)计算。

$$P(X|C_i)=\prod_{k=1}^{n}P(X_k|C_i) \quad (3)$$

其中，$P(X_k|C_i)$即表示条件 X_k 对分类结果为 C_i 的影响程度。对于任意的输入向量 X，式(3)计算训练数据中分类为 C_i 的情况下，输入向量中元素出现的概率。当 X 中的某个元素值在分类为 C_i 的训练数据记录中出现的次数越多，说明该元素值对分类结果为 C_i 的影响程度越高。最后再综合各个元素的概率计算分类为 C_i 的概率。因此，对于任意的背景变化，贝叶斯方法均能合理地进行分类。

本文实验中所使用的数据集为小样本数据。因此，在计算 $P(X|C_i)$时，肯定存在某些 $P(X_k|C_i)$的概率值为零的情况，而这会导致整个 $P(X|C_i)$的概率计算为零。假设当没有这个零概率时，实验可能得到一个表明 X 属于 C_i 类的高概率，而一个零概率将完全消除 C_i 上其他概率的影响。因此为了降低零概率值对分类结果的影响，需要对这些概率值为零的项进行校准，使得需要的分类结果对 C_i 记录中属性 X_k 不同取值的计数加 1 造成的估计概率的变化可以忽略不计，也称为拉普拉斯校准。然而，数据集中组织生成的变化表记录数较少，校准会使原本值为零的概率出现较大的变化。因此可以在出现零概率值时，将计数加上一个较小的数(例如 0.01)使估计概率的变化可以忽略。

2.2 贝叶斯方法预测恐怖行为

使用贝叶斯分类算法在变化表上预测恐怖行为。对于变化向量 X 及分类结果 C_i，算法在变化表的每一条记录中搜索变化表的分类结果。如果分类结果为 C_i，则检查变化表的属性是否与变化向量 X 中对应的属性值相等。最终统计出变化表中分类结果为 C_i 且各个属性等于变化向量 X 中对应的属性值的概率，计算 $P(X|C_i)$。算法伪代码如下所示。

```
输入:查询变化向量 X
输出:分类结果
1. C 为包含行为 Aj 的各个值的数组;
2. n 为背景属性的数目;
3. prob 数组,长度与 C 相同,初始化为 1;
4. for each element Ci in C do
5.     size=0;
6.     count 为统计数组,包含 n+1 个元素,赋值为 0;
7.     for each row CT in CT(g,Aj) do
8.         if CT[n+2]=Ci then
9.             size++;
10.            for i from 1 to n+1 do
11.                if CT[i] matches Xi then
12.                    count[i]++;
13.                end
14.            end
15.        end
16.    end
17.    for each element in count do
18.        if count[index]=0 then
19.            //拉普拉斯校准
20.            count[index]=0.01;
21.        end
22.        prob[i]=prob[i] * count[index]/size;
23.    end
24.    prob[i]=prob[i] * P(Ci);
25. end
26. return Ci 当 prob[i]在 prob 数组中最大时。
```

伪代码中，步骤 4－步骤 25 用于计算在背景向量 X 的情况下，分类为 C_i 的概率，即 $P(C_i|X)$。步骤 7－步骤 16 扫描变化表 $CT(g,A_j)$，查找变化表中分类为 C_i 的行中，背景属性的值为 X_k 的次数用以计算 $P(X_k|C_i)$。步骤 17－步骤 25 实现拉普拉斯校准并完成 $P(X|C_i)$的计算。步骤 24 根据贝叶斯定理计算 $P(C_i|X)$。

令 m 为数据集中记录数，n 为背景属性的个数，c 为行为属性 $A_j\in AS(g)$中的分类数目。基于贝叶斯方法预测恐怖行为时，只需要在整个训练集上计算 $P(X_k|C_i)$，即步骤 5－15 的三重循环，故其时间复杂度为 $O(cmn)$。

2.3 加权贝叶斯预测恐怖行为

背景的改变对行为的影响不是转瞬即逝的，可能会在一定的时间内对行为产生持续的影响。因此，需要在不同时间滞差情况下预测组织的行为，并将不同时间滞差下预测的结果进行综合评估。加权贝叶斯预测方法分为训练和预测两步。其中，训练过程根据训练变化表中的数据预测行为序列，计算预测的行为序列与真实行为的相关性。将得到的各步相关性进行归一化，作为最后预测过程中不同时间滞差预测结果的系数。

令 $X_i^h=(x_i^h,y_{i+1})$为 h-变化表 $CT^h(g,A_j)$中的第 i 时刻数据($2\leqslant i\leqslant m-h$)，$y_{i+1}$为时间段 $i+1$ 的行为值。令 $Y^h=(y_{h+2},\cdots,y_m)$为组织从时间 $h+2$ 到 m 的实际行为时间序列，$PY^h=(y'_{h+2},\cdots,y_m')$为在 h-变化表 $CT^h(g,A_j)$上使用贝叶斯算法得到的预测行为序列，即有

$$PY_i^h=Bayes\_forecast(x_i^h) \quad (4)$$

那么 h 步预测的权重系数由式(5)、式(6)得到：

$$r_h=\sum(Y^h{}_i-\overline{Y^h})(PY^h{}_i-\overline{PY^h})/\sqrt{\sum(Y^h{}_i-\overline{Y^h})^2\sum(PY^h{}_i-\overline{PY^h})^2} \quad (5)$$

$$w_h=|r_h|/\sum|r_k| \quad (6)$$

分析以上公式可以得知，如果预测行为序列 PY^h 与真实行为序列 Y^h 越相似，则相关系数 $|r_h|$越接近于 1，说明 h 步的变化表具有较好的预测效果，那么在预测过程中应具有更高的权重。

预测的目标函数为：

$$\max_h\{\sum_h w_h * \frac{P^h(C_i|X)}{\sum_k P^h(C_k|X)}\} \quad (7)$$

其中，$P^h(C_i|X)$表示在 h-变化表中计算 $P(C|X)$的概率。

对于每种行为而言，计算在各步变化表上使用贝叶斯分类预测得到该行为的概率在所有分类结果概率中的比重之和，选择使结果数值最大化的行为作为分类结果。

上述算法的步骤为：

1)构建 h-变化表 $CT^h(g,A_j)$。

2)h-变化表 $CT^h(g,A_j)$的每一行的变化数据使用贝叶斯方法预测，得到预测的行为序列 PY^h。

3)根据式(5)计算预测行为序列 PY^h 与真实行为序列 Y^h 的相关系数 r_h。

4)判断是否是最大步长？如果不是，则跳到步骤 1)计算

下一个步长的相关系数；否则，进行下一步。

5)根据式(6)将得到的相关系数 $r_1, r_2, \cdots, r_h$ 归一化，得到预测权重系数 $w_1, w_2, \cdots, w_h$。

若需要预测在输入为 X 时的组织行为类别，只需根据式(7)计算各个行为类别的得分，输出得分最高的行为类别。

从理论上来说，多步变化表预测模型中的步长越大时，模型的预测效果更佳。但是 MAROB 数据集中数据的时间周期为 1 年，根据实际经验，步长 h 的值建议小于 5，即每个背景因素的影响在 5 年以后几乎为零。最佳的 h 值与组织的活动规律、数据集的时间周期以及数据集的规模等因素相关。因此，较优的 h 值仍需要通过对组织进行实验观察获得。

贝叶斯分类预测算法的时间复杂度为 $O(cmn)$，则 h 步的加权贝叶斯预测算法训练和预测过程的算法时间复杂度均为 $O(hcmn)$。相比之下，CAPE 算法总的时间复杂度为 $O(2^n km^2 n^2)$(k 为 CONVEX 算法中的参数，表示 k 个最近背景向量)，贝叶斯算法的时间复杂度明显优于 CAPE 算法。

3 实验与分析

在实验中，以 MAROB 数据集中的 Hamas(组织编号 6660312)等恐怖组织的数据作为实验数据，以准确率与运行时间为测试指标，并以 CAPE 算法及 SitCAST+CONVEX 算法作为本文所提 Bayes 方法的比较对象。实验运行环境是 Core i3 M370，2.4GHz，2048MB，Microsoft Windows7，算法使用 Matlab 编程实现。实验属性数据中的缺失值使用其样本的均值填补。

对 Hamas 恐怖组织的 15 种行为进行预测的测试结果如表 4 和图 1 所示。其中，Bayes 1-CT 表示单独使用 1 步变化表预测得到的准确度。

表 4 平均预测准确率及算法运行时间

预测算法	Precision(%)	Time(s)
Bayes 1-CT	94.58	0.0218
CAPE	80.42	2.9174
SitCAST+CONVEX	65.83	3.8789

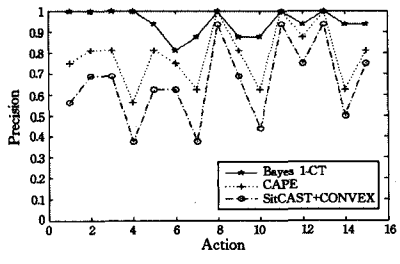

图 1 不同算法的预测结果

在预测精度上，从表 4 可见，Bayes 算法的平均预测精度为 94.6%，明显高于 CAPE 算法的 80.4% 和 SitCAST+CONVEX 算法的 65.8%。从图 1 也可看出，Bayes 算法在每一种恐怖行为预测上，其预测精度都不低于其它两种，3 种算法的预测精度与表 4 中的平均预测精度一致。究其原因是因为贝叶斯定理是一种把类的先验知识和从数据集中收集的类的后验知识相结合的统计学方法，权衡了各个条件对分类结果的影响，因而相对于其它分类算法，贝叶斯分类算法具有较好的预测准确率。而 CAPE 算法准确率偏低的原因主要是由 SitCAST 算法造成的，该算法利用各个背景属性单独使用线性回归、非线性回归等方法预测属性值，背景预测结果的误差是它的各个属性值预测误差的叠加。

在执行时间上，从表 4 可以看出 Bayes 算法的执行时间明显低于 CAPE 算法及 SitCAST+CONVEX 算法，执行速度比其他两类算法提高了两个数量级。对于每一个输入向量，Bayes 算法只需扫描 c(类划分的数量)遍变化表，统计输入向量在每种分类情况下出现的概率，查找最大概率的分类，将其作为分类预测结果。而一般情况下，CAPE 大多使用 SitCAST+CONVEX 算法预测：首先需要根据背景中每一项属性以往的值预测下一时刻该属性的值，将该值向上取整的值和向下取整的值作为该属性的可能值，生成所有可能的背景属性集合，然后对每一个可能的背景属性集合使用 CONVEX 算法预测行为，这是指数级时间复杂度的过程。因此，Bayes 算法的时间复杂度明显低于 CAPE 算法。

对 Hamas 恐怖组织的 15 种行为使用多步变化表及多变化表加权预测的测试结果如表 5 所列。在不同步长的变化表上使用贝叶斯预测时，预测准确率呈现先增加后减小的趋势，恰好符合背景变化在刚开始时产生的影响慢慢加强，而随着背景变化的影响达到峰值后开始减弱的预期。此外，使用加权贝叶斯预测比使用单个变化表预测在准确率上有略微的提升。

表 5 不同变化表上的预测效果及算法运行时间

变化表	Precision(%)	Time(s)
1-CT	94.58	0.0218
2-CT	94.67	0.0286
3-CT	96.67	0.0240
4-CT	95.38	0.0161
2 步加权	95.47	0.0517
3 步加权	96.85	0.0763
4 步加权	96.51	0.0919

虽然从理论上来说，多步变化表预测模型比单步变化表预测模型的预测效果更佳，但是限于数据集的规模，最佳的步数 h 仍需要通过大量的实验观察获得。

为了验证贝叶斯方法预测恐怖组织行为算法对于其他组织有效，选择背景属性变化较多的 9 个组织的数据进行实验，其测试结果如表 6 所列。从表 6 中可以看到，贝叶斯方法作为恐怖行为预测方法对于实验中组织的行为预测具有较好的预测效果，其预测准确率高于其他两种方法。

表 6 使用不同方法预测组织行为的准确率

名称	Bayes	CAPE	SitCAST+CONVEX
Hamas	0.9458	0.8042	0.6583
Komala	0.8261	0.6957	0.6522
KSDP	0.8235	0.4706	0.3529
ICP	0.9348	0.8478	0.7609
Dawa	0.8116	0.6087	0.5507
PLO1	0.9565	0.8913	0.8261
PLO2	0.8696	0.8571	0.8075
DFLP	0.8860	0.8421	0.7807
PFLP	0.9013	0.8158	0.7632

结束语 本文指出了大部分传统方法的缺陷，并介绍了预测组织行为变化的 CAPE 算法。通过分析 CAPE 模型中存在的预测准确率较低、时间复杂度高且算法流程复杂等缺点的原因，提出了一种在改进的变化表上，使用贝叶斯方法预测恐怖行为的算法。实验表明，基于贝叶斯方法预测恐怖行为算法的准确率优于 CAPE 模型，其耗时降低了两个数量级。此外，考虑到背景的改变会在时间序列上对行为产生持

续的影响，因此建立了在不同影响滞差情况下，利用加权贝叶斯方法预测组织行为的模型。

参 考 文 献

[1] Serra E, Subrahmanian V S. A Survey of Quantitative Models of Terror Group Behavior and an Analysis of Strategic Disclosure of Behavioral Models[J]. IEEE Transactions on Computational Social Systems, 2014, 1(1): 66-88

[2] Subrahmanian V S. Handbook of Computaional Approaches to Counterterrorism[M]. New York: Springer Press, 2013: 99-268

[3] Xue An-rong, Wang Wei, Zhang Ming-cai. Terrorist Organization Behavior Prediction Algorithm Based on Context Subspace [M]// Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2011: 332-345

[4] Li Xiao-chen, Mao Wen-ji, Zeng D, et al. Performance evaluation of machine learning methods in cultural modeling[J]. Journal of Computer Science and Technology, 2009, 24(6): 1010-1017

[5] Subrahmanian V S. Cultural Modeling in real time[J]. Science, 2007, 317(5844): 1509-1510

[6] Khuller S, Martinez V, Nau D, et al. Finding Most Probable Worlds of Probabilistic Logic Programs[C]// 1st International Conference on Scalable Uncertainty Management. Washington DC: IEEE Press, 2007: 45-57

[7] Subrahmanian V S, Albanese M, Martinez M V, et al. CARA: a cultural adversarial reasoning architecture[J]. IEEE Intel Syst, 2007, 22(2): 12-16

[8] Martinez M V, Simari G I, Sliva A, et al. CONVEX: context vectors as a similarity-based paradigm for forecasting group behaviors[J]. IEEE Intel Syst, 2008, 23(4): 51-57

[9] Martinez M V, Sliva A, Simari G I, et al. CAPE: automatically predicting changes in group behavior[C]// Mathematical Methods in Counterterrorism. Springer Vienna, 2009: 253-269

[10] Victor A, Pate A, Wilkenfeld J. Minorities at risk organization behavior data and codebook version 9[EB/OL]. http://www.cidcm.umd.edu/mar

[11] Guo Hong-yu. Research on weighted feature selection algorithm based on information entropy theory[J]. Computer Engineering and Applications, 2013, 49(10): 140-146(in Chinese)
郭红钰. 基于信息熵理论的特征权重算法研究[J]. 计算机工程与应用, 2013, 49(10): 140-146

[12] Liu Hua-wen. Research on feature selection algorithm based on information entropy[D]. Changchun: Jilin University, 2010(in Chinese)
刘华文. 基于信息熵的特征选择算法研究[D]. 长春: 吉林大学, 2010

[13] He Qing, Li Ning, Luo Wen-juan, et al. Summary of machine learning algorithms in large data[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(4): 327-336(in Chinese)
何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能, 2014, 27(4): 327-336

[14] Han Jia-wei, Kamber M. 数据挖掘：概念与技术(3rd ed)[M]. 范明, 等译. 北京: 机械工业出版社, 2012

(上接第 124 页)

$\mu_{w,d}^{k}$，故它们的空间复杂度都为 $O(K*(D+W))$。而 Infvoc-LDA 对于每个主题分布中的单词需要保存两个参数，空间复杂度为 $O(K*(D+2*W))$。由于都是在线算法，可以通过调整 *minibatch* 来控制 dvOBP 的空间复杂度，但是在 Infvoc-LDA 中有内存隐患，因为主题单词矩阵是全局变量，需要常驻内存。

结束语 本文首先提出了传统的 LDA 算法中固定词汇表的问题，然后提出了一种新的基于动态词汇表的 LDA 算法 dvOBP，并给出了详细的推导过程。之后将其与其他固定词汇表的传统 LDA 算法以及同样基于动态词汇表的 Infvoc-LDA 算法在混淆度、互信息指数以及收敛时间上进行比较，可以发现，dvOBP 在词汇表上的解决方案虽然存在着时间复杂度比传统的 LDA 算法偏高的问题，但是确实可以解决实际问题中固定词汇表所造成的问题，在混淆度、互信息指数上优于传统的 LDA 算法。

参 考 文 献

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3(1): 993-1022

[2] Zeng J, Cheung W K, Liu J. Learning topic models by belief propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(5): 1121-1134

[3] Heinrich G. Parameter estimation for text analysis[R]. Technical Report, 2005

[4] Sethuraman J. A constructive definition of Dirichlet priors[R]. Florida State Univ Tallahassee Dept of Statistics, 1991

[5] Zhai K, Boyd-Graber J. Online Latent Dirichlet Allocation with Infinite Vocabulary[C]// Proceedings of The 30th International Conference on Machine Learning. 2013: 561-569

[6] Mimno D, Hoffman M, Blei D. Sparse stochastic inference for latent Dirichlet allocation[J]. arXiv, 2012(3): 362-365

[7] Newman S K D, Cavedon L. External evaluation of topic models [C]// Australasian Document Computing Symposium. 2012: 11-18

[8] Hoffman A F M, Blei D. Online inference of topics with latent dirichlet allocation[C]// NIPS. 2010: 856-864

[9] Yao L, Mimno D, McCallum A. Efficient methods for topic model inference on streaming document collections[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009: 937-946

[10] Zeng J, Liu Z Q, Cao X Q. Online belief propagation for topic modeling[J]. arXiv preprint arXiv: 1210. 2179, 2012

[11] Ishwaran H, Zarepour M. Dirichlet prior sieves in finite normal mixtures[J]. Statistica Sinica, 2002, 12(3): 941-963

[12] Mei S Y, Wang F, Zhou S G. Dirichlet process mixture model, extensions and appication[J]. Chin Sci Bull, 2012, 57(34): 3243-3257(in Chinese)
梅素玉, 王飞, 周水庚. 狄利克雷过程混合模型、扩展模型及应用[J]. 科学通报, 2012, 57(34): 3243-3257

[13] Gong Sheng-rong, Ye Yun, Liu Chun-ping, et al. Topic Tracking Based on Online Belief Propagation[J]. Chinese Journal of Computers, 2015, 38(2): 249-260(in Chinese)
龚声蓉, 叶芸, 刘纯平, 等. 基于在线消息传递的主题追踪方法[J]. 计算机学报, 2015, 38(2): 249-260