计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(9):200−204 [doi: 10.15888/j.cnki.csa.005995]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

适用于手机取证数据的权重计算方法①

肖程望, 卢　军, 余力耕, 张　弛

(武汉邮电科学研究院, 武汉 430074)

摘　要: 传统分类系统往往选择朴素贝叶斯算法作为分类算法, 在研究过程中, 发现朴素贝叶斯模型(NBC)具有以下前提条件: 所有特征互不影响, 并且特征属性的权重为 1. 研究后发现并非如此, 在对数据进行分类时, 有的特征可能对分类结果的影响较大, 有的可能对结果影响较小. 为了对算法进行优化, 需要给不同的特征附上不同的权重值, 才能更加客观的获得分类结果. 本文研究了两种传统的基于属性权重的计算方法, 同时考虑到手机取证数据的特点, 提出两种适应于手机取证数据的改进权重计算方法, 并对其改进原理进行研究, 比较改进后的权重计算方法与传统的权重计算方法, 在使用相同分类算法与相同数据的情况下, 对分类结果的影响.

关键词: 手机取证; 权重计算; 分类算法

引用格式: 肖程望,卢军,余力耕,张弛.适用于手机取证数据的权重计算方法.计算机系统应用,2017,26(9):200–204. http://www.c-s-a.org.cn/1003-3254/5995.html

Weight Calculation Method for Mobile Phone Forensics Data

XIAO Cheng-Wang, LU Jun, YU Li-Geng, ZHANG Chi

(Wuhan Research Institute of Posts and Telecommunications, Wuhan 430074, China)

Abstract: The traditional classification system often chooses the Naive Bayesian algorithm as the classification algorithm. In the course of the study, we find that the Naive Bayesian model(NBC) has the following conditions: all the characteristics do not mutually influence each other, and the feature attribute weights is 1. But we find that is not the case after a study. In the classification of data, some features may have a greater impact on the classification results, while some may have little impact. In order to optimize the algorithm, we need to attach different weights to different features, so as to obtain the classification results more objectively. This paper studies two kinds of calculation methods of attributing weight based on the traditional algorithm. At the same time, considering the characteristics of mobile phone forensic data, it proposes the calculation method of two kinds of improved weight suitable for mobile phone forensic data. This paper researches the improvement principle of research, compares the improved calculation method of weight with the traditional calculation method in their impacts on the classification results using the same classification algorithm with the same data.

Key words: mobile phone forensics; term weighting; method of forensics

随着信息技术飞速发展, 智能手机已然成为人们日常生活中必不可少的工具. 由于智能手机变得越来越流行, 利用其进行犯罪活动的行为也越发的频繁. 这一愈发突出的现象, 要求我们对手机取证方面进行相关研究, 从中发现重点信息和需要重点关注的对象.

目前, 基于手机取证数据的分析算法研究主要集中在两个方面: 一方面针对手机内信息进行亲密度分析, 例如分析手机使用者与通讯录内各联系人的亲密度关系, 主要利用各种分类算法, 并对现有的分类算法进行改进, 其中改进的方向主要集中在权重计算方法

① 收稿时间: 2017-01-03; 采用时间: 2017-02-17

的改进与优化上[1]; 另一方面针对手机内信息进行关联规则挖掘, 例如通过 Apriori、FP-tree 等关联规则挖掘算法分析团伙每个人手机中的信息, 分析出犯罪团伙中的主要人物或关键人物[2]. 本文提出适应于手机取证数据的改进权重计算方法, 基于属性频率和变异系数的权重计算方法, 并对其改进原理进行研究. 最后通过实验对权重计算准确性进行验证, 证明本文提出的权重计算方法的准确性.

1 算法分析

1.1 传统权重计算方法分析

传统分类系统往往选择朴素贝叶斯算法作为分类算法, 在研究过程中, 发现朴素贝叶斯模型(NBC)具有以下前提条件: 所有特征互不影响, 并且特征属性的权重为 1. 研究后发现并非如此, 在对数据进行分类时, 有的特征可能对分类结果的影响较大, 有的可能对结果影响较小[3]. 为了对算法进行优化, 需要给不同的特征附上不同的权重值, 才能更加客观的获得分类结果, 下面分析两种传统的权重计算方法.

1.1.1 基于属性频率的权重计算方法

在对测试数据进行分类时, 向量 C 为属性条件集, 其中分类样本有 n 个可能出现的属性, 当有 m 个训练样本时, 可以组成 $m*n$ 阶的矩阵, 记为 $M=(M_n)_{m*n}$. 由属性频率定义可知, 属性在此矩阵中出现的频率越高, 那么重要性就越高, 对分类结果的影响也就越大[4]. 式(1)表示了属性 a 所对应的属性频率, 应用在针对手机取证数据中的属性权重分析时, 需要依靠属性的重要性对频率值进行重新分配, 得出式(2)的基于属性频率的权值计算方法.

$$\gamma(a)=\{m|a\in m,\forall m\in M\} \tag{1}$$

$$w_a=\frac{\gamma(a)}{\frac{1}{n}\sum_{i=1}^{n}\gamma(a_i)} \tag{2}$$

为验证算法的有效性, 基于 UCI 数据库, 利用属性频率权重计算方法对不同数据集的属性进行加权后, 将不同的权重值带入三种不同的分类算法中进行验证, 得出了如表 1 所示数据.

1.1.2 基于相关系数的权重计算方法

相关关系是一种非确定性关系, 相关系数是研究变量之间相互关联程度的量. 应用在分类系统中时, 设测试样本具有 n 个条件属性与 1 个决策属性, 可用 $M_i(i=1, 2, 3\ldots n)$和 N 表示, 则可得第 i 个特征属性所对应的权重系数[5], 如式(3)所示:

$$W_i=\left|\rho_{Mi,N}\right|=\left|\frac{Cov(Mi,N)}{\sqrt{D(Mi)D(N)}}\right| \tag{3}$$

表 1 基于属性频率的权值计算方法实验结果

数据集	属性	类别	训练集	NB(%)	KNN(%)	SVM(%)
Cleve	10	2	396	82.43	82.92	82.31
Australian	14	2	890	78.63	80.26	79.96
Cars	7	4	1800	86.96	83.51	86.37
Vehicle	18	4	848	63.38	63.42	62.89
Iris	4	3	150	92.71	92.66	92.51
Segment	7	18	2300	93.26	92.08	92.66
Glass	9	7	214	73.77	74.62	71.35
均值				81.59	81.35	81.15

式中 $Cov(Mi, N)$的计算方法为 $E(Mi\ N)$-$E(Mi)E(N)$, 由此式得出的特征属性所对应的权重系数为一个介于 0 到 1 之间的常数, 当 W_i 为 0 时, 表示此特征属性对分类结果没有影响, 当 W_i 为 1 时, 表明此特征属性对分类结果的影响较大, 呈线性关系, 并且当 W_i 越趋近与 1, 影响越大, 当 W_i 趋近于 0, 特征属性对分类结果的影响较小[6]. 由此可知, W_i 表示特征属性 M 与决策属性之间的相关性程度, 可应用在手机取证分析系统中作为加权系数优化分类结果. 应用上一节中所用数据库, 利用相关系数权重计算方法对不同数据集的属性进行加权后, 将不同的权重值带入三种不同的分类算法中进行验证, 结果如表 2 所示.

表 2 基于相关系数的权值计算方法实验结果

数据集	属性	类别	训练集	NB(%)	KNN(%)	SVM(%)
Cleve	10	2	396	91.17	90.42	92.11
Australian	14	2	890	80.49	80.73	80.42
Cars	7	4	1800	91.21	91.16	92.05
Vehicle	18	4	848	76.72	77.21	76.97
Iris	4	3	150	97.83	96.93	97.26
Segment	7	18	2300	98.89	98.90	99.12
Glass	9	7	214	93.20	93.35	94.12
均值				89.93	89.81	90.29

1.2 适用于手机取证数据的权重计算方法

考虑到手机中提取出的数据具有, 样本小、属性多的特点, 以上两种权重计算方法主要是针对数据量较大的情况, 且都是从正向考虑特征属性对分类结果的影响, 应用在手机取证中计算结果可能会不太准确. 并且属性权重计算是分类算法中十分重要的一部分, 下面对两种改进后的权重计算方法进行研究, 改进的理论依据在于, 考虑到的信息量的熵值越大, 其携带的信息越多, 对分类结果的影响也就越大[7].

1.2.1 基于变异系数的权重计算方法

变异系数法(Coefficient of variation method) 是一种客观赋权的方法, 在很多场合都有所应用, 理论依据为利用各个特征项所包含的信息大小, 来决定各个特征项的权重值[8]. 因为在评价一类事物时, 相互间差别越大的特征项越能表达这类事物的不同之处, 更能反映相互之间的差别. 针对手机取证中数据的特点, 引入变异系数作为计算特征属性权重的一种方法.

将各属性视为随机变量 M_i, 任意随机变量 M_i 的标准差与平均数的比值称为对应的变异系数, 记为 CV_i, 可得到各属性对应权重. 在评价通讯录内联系人的亲密度时, 有多种评价标准, 例如: 通话次数、通话时长、短信次数、短信中关键字词的出现频率、邮件联系次数等等. 由于各个指标的量纲不同, 不能直接拿来进行比较, 需进行归一化处理. 考虑到概率计算后都为大于 0 小于 1 的数, 需进行比例逆运算, 然后得到各个指标的权重系数. 计算过程如下:

(1) 对训练数据进行分析训练数据, 分别计算特征属性的平均数和标准差;

(2) 按式(4)计算出变异系数(均值与标准差的比值);

$$C = \frac{\mu}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}} \tag{4}$$

(3) 将特征属性所对应的变异系数相加求和, 然后进行归一化处理;

(4) 通过比例逆运算计算特征属性所对应的权重系数.

通过变异系数法得出权重系数, 量化了不同特征属性对分类结果的影响, 可应用在加权贝叶斯分类模型中, 对分类算法进行优化.

1.2.2 基于属性频率和变异系数的权重计算方法

基于属性频率的权重计算方法考虑到正向类别对分类结果的影响, 例如某属性出现频率较高时, 对分类类别的影响较大, 然而, 未能考虑到信息量的熵对分类结果的影响, 比如某个属性虽然出现频率较低, 但对分类结果具有决定性的影响. 同时, 基于变异系数的权重计算方法考虑到, 相互间差别越大的特征项越能表达事物之间的不同之处, 更能反映属性之间的差距. 举例说明上述问题: 从数据库中随机选取两个特征项, 利用两种权重计算方法进行测试, 结果如表 3 所示.

表 3 两种权重计算方法简单比较

特征项	属性出现次数	变异系数	基于属性频率	基于属性频率和变异系数
特征项1	3	0.781	3.353	10.760
特征项2	3	0.574	3.353	7.286

从上表可知, 特征项 1 和特征项 2 有相同的属性频率, 但所包含的信息量有区别, 单从属性频率方面考虑不能得到准确的属性权重, 赋予特征项 1 和特征项 2 相同的权重是不合适的. 只有同时考虑到属性频率与变异系数, 得出的属性权重才符合实际情况. 因此, 综合考虑两种算法的优点与缺陷, 引入基于属性频率和变异系数的权重计算方法, 计算公式如式(5)所示:

$$\rho_i = \frac{\gamma(a)}{\frac{1}{n}\sum_{i=1}^{n}\gamma(a_i)} * \frac{\mu}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}} \tag{5}$$

该方法理论依据为: 在使用变异系数计算方法进行分析前, 先从全局的角度使用基于属性频率的权重计算方法对各属性的进行权重评估, 然后通过变异系数计算方法对其进行改进, 这样便可以削弱特征属性的频率对各类别的代表性. 这样, 在表 3 的例子中, 特征项 1 比特征项 2 的权重要高, 是因为特征项 1 所携带的信息量更多, 增强了其类别的代表性.

2 算法设计与实现

本节主要实现在上一节中提出的四种权重计算方法, 包括基于属性频率的权重计算、基于相关系数的权重计算和新提出的两种更适用于手机取证数据的基于变异系数的权重计算和基于属性频率和变异系数的权重计算方法, 并通过实验进行比较, 在得出各个算法的权重计算结果后, 进行归一化的处理. 图 1 展示特征属性权重计算模块的具体流程图.

由于四种特征属性权重计算方法的流程几乎一样, 区别仅仅是计算公式不同, 以下以基于属性频率的权重计算方法为例, 图 2 为此方法的伪代码实现.

3 实验结果分析

本节主要对上一节中设计的分类器进行测试, 通过使用四种特征属性权重计算方法对 10 部真实手机内数据的各个特征属性进行加权, 包括通话时长/通话次数、短信联系频率、微信聊天频率与短信关键词出现频率等, 然后带入三种分类算法中进行分析, 同时利

用经典的 F1 评价方法对分类性能进行评估, 如式(6)所示.

$$F1 = \frac{2P*R}{P+R} \quad (6)$$

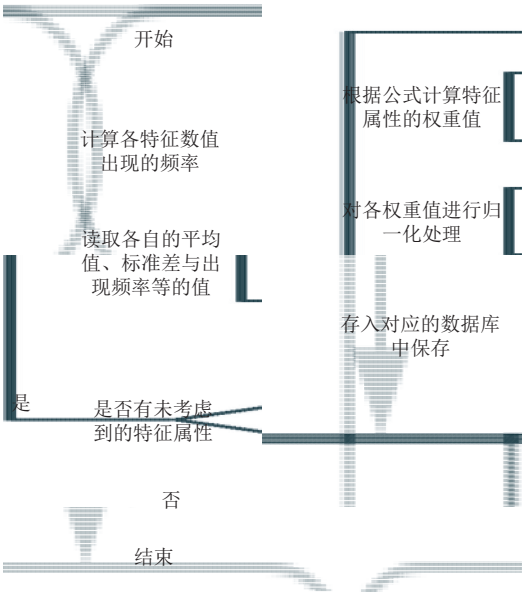

图 1 特征属性的权重计算流程图

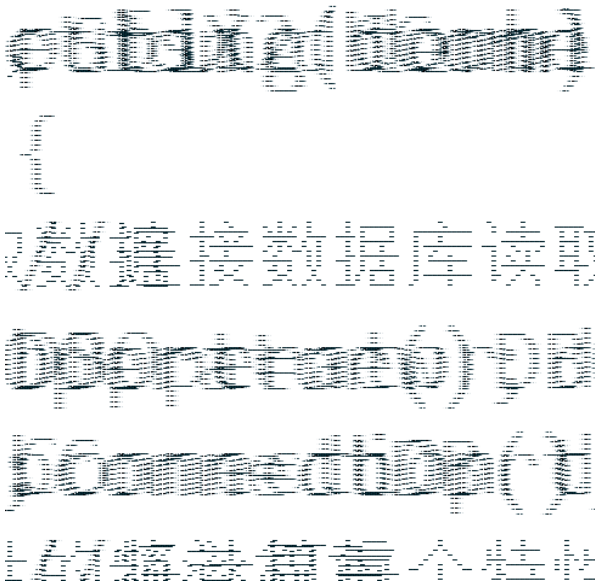

图 2 特征权重计算伪代码实现

从公式中可发现, F1 评价方法综合考虑了正确率与召回率两个参数, 当 F1 值较高时, 分类效果较好, 实验结果如表 4 所示.

从实验数据中可知:

(1) 在基于属性频率权重计算方法、相关系数权重计算方法、基于变异系数权重计算方法与基于属性频率和变异系数权重计算方法中, 基于属性频率和变异系数的权重计算方法准确性最高;

(2) 根据手机取证的数据具有小样本、多属性的特点, 选择基于变异系数的权重计算方法和改进后的基于属性频率和变异系数权重计算方法应用在手机取证中, 比传统的权重计算方法表现得更好;

(3) 在传统计算机取证中应用较好的基于相关系数和基于属性频率的权重计算方法 F1 值在 0.8 左右, 然而对于手机取证的数据来说, 更加关注召回率, 在这方面, 这两种方法均没有基于变异系数和改进的基于属性频率和变异系数的方法效果好;

(4) 在传统取证过程中运用较好的 SVM 和 KNN 分类算法, 在手机取证中分类效果虽然整体来看效果较好, 但是优势有限, 在某些特征权重计算的优化下, 甚至不如 NB 分类算法.

表 4 实验结果

权重计算方法	分类算法	Precision	Recall	F1-Measure
基于属性频率	Naive Bayes	0.713	0.853	0.777
	SVM	0.652	0.916	0.762
	KNN	0.727	0.832	0.776
基于相关系数	Naive Bayes	0.882	0.730	0.799
	SVM	0.903	0.775	0.834
	KNN	0.897	0.721	0.800
基于变异系数	Naive Bayes	0.871	0.845	0.858
	SVM	0.839	0.870	0.854
	KNN	0.871	0.855	0.863
基于属性频率和变异系数	Naive Bayes	0.923	0.887	0.905
	SVM	0.896	0.910	0.902
	KNN	0.907	0.886	0.896

4 结语

本文着重对应用在分类系统中的几种不同权重计算方法进行了研究, 并且基于手机取证中的数据特征, 提出了更适用于手机取证数据的基于属性频率和变异系数的权重计算方法, 使用测试数据对不同的方法进行了测试比较, 分析算法的优化原理, 并且设计算法的实现步骤. 通过对实验结果的分析, 证明本文提出的权重计算方法较为优秀和稳定.

同时, 存在以下不足之处, 需要进一步的研究:

(1) 本文对各种不同分析方法的测试主要通过分类精度进行比较, 而测试数据的亲密度是人为评定的, 有一定的误差, 后续需要收集更多的数据进行分析, 以便得到更精确的数据;

(2) 还可对本文提出的算法继续进行优化, 例如利用本文提出的方法和更多权重计算方法进行组合.

下一步的研究方向主要是考虑新的变异系数度量方法以便更进一步的提高分类性能, 以及考虑各属性的其他特征以及各属性间的相关性.

参考文献

1 周喜. 基于粗糙集的加权朴素贝叶斯分类算法研究[硕士学位论文]. 长沙: 长沙理工大学, 2013.

2 贾娴, 刘培玉, 公伟. 基于改进属性加权的朴素贝叶斯入侵取证研究. 计算机工程与应用, 2013, 49(7): 81–84.

3 刘磊, 陈兴蜀, 尹学渊, 等. 基于特征加权朴素贝叶斯分类算法的网络用户识别. 计算机应用, 2011, 31(12): 3268–3270.

4 徐光美, 刘宏哲, 张敬尊. 基于特征加权的多关系朴素贝叶斯分类模型. 计算机科学, 2014, 41(10): 283–285. [doi: 10.11896/j.issn.1002-137X.2014.10.059]

5 饶丽丽, 刘雄辉, 张东站. 基于特征相关的改进加权朴素贝叶斯分类算法. 厦门大学学报(自然科学版), 2012, 51(4): 682–685.

6 Huang LS, Moshchuk A, Wang HJ, *et al*. Clickjacking: Attacks and defenses. Proc. the 21st Usenix Conference on Security Symposium. Bellevue, WA, USA. 2012. 22.

7 Yao YY, Zhao Y. Attribute reduction in decision-theoretic rough set models. Information Sciences, 2008, 178(17): 3356–3373. [doi: 10.1016/j.ins.2008.05.010]

8 杨敏, 贺兴时. 基于改进的加权贝叶斯分类算法在空间数据中的应用. 价值工程, 2012, 31(36): 201–203. [doi: 10.3969/j.issn.1006-4311.2012.36.101]

9 李卫平, 杨杰, 王钢. 比例逆权重 kNN 算法及其流处理应用. 计算机工程与设计, 2015, 36(12): 3355–3358.

10 秦锋, 任诗流, 程泽凯, 等. 基于属性加权的朴素贝叶斯分类算法. 计算机工程与应用, 2008, 44(6): 107–109.

11 程克非, 张聪. 基于特征加权的朴素贝叶斯分类器. 计算机仿真, 2006, 23(10): 92–94, 150. [doi: 10.3969/j.issn.1006-9348.2006.10.024]

12 张明卫, 王波, 张斌, 等. 基于相关系数的加权朴素贝叶斯分类算法. 东北大学学报(自然科学版), 2008, 29(7): 952–955.

13 鲁明羽, 李凡, 庞淑英, 等. 基于权值调整的文本分类改进方法. 清华大学学报(自然科学版), 2003, 43(4): 513–515.

14 陈晓琳, 姬波, 叶阳东. 一种基于 ReliefF 特征加权的 R-NIC 算法. 计算机工程, 2015, 41(4): 161–165.

15 王小丽, 远俊红. 基于加权朴素贝叶斯分类法的成绩预测模型. 电子技术与软件工程, 2013, (19): 225–226.