(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110515836 A
(43)申请公布日 2019.11.29

(21)申请号 201910700517.4

(22)申请日 2019.07.31

(71)申请人 杭州电子科技大学
地址 310018 浙江省杭州市下沙高教园区2号大街

(72)发明人 王兴起 王赛 魏丹 陈滨 邵艳利 王大全

(74)专利代理机构 杭州君度专利代理事务所（特殊普通合伙）33240
代理人 杨舟涛

(51)Int.Cl.
*G06F 11/36*(2006.01)
*G06K 9/62*(2006.01)

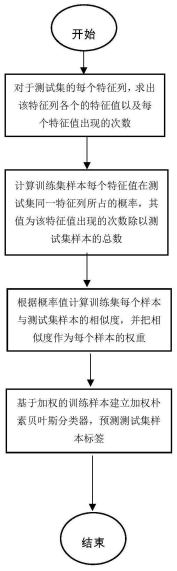
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种面向软件缺陷预测的加权朴素贝叶斯方法

(57)摘要

本发明公开了一种面向软件缺陷预测的加权朴素贝叶斯方法，现有技术中朴素贝叶斯方法没有考虑到训练集数据与测试集数据之间的相似性对预测结果的影响。已有的朴素贝叶斯的改进方法没有考虑到某一特征值在该特征属性中所占的概率大小的影响，所计算的相似性不够准确，因此对样本权重计算不够精确，会影响分类效果的准确性。本发明所提出的一种面向软件缺陷预测的加权朴素贝叶斯方法能够根据训练集样本与测试集样本的相似度为训练集样本加权，并且能够考虑到某一特征值在该特征属性中所占的概率大小的影响。因此，本发明可以提高朴素贝叶斯的预测性能。




CN 110515836 A

1.一种面向软件缺陷预测的加权朴素贝叶斯方法,其特征在于包括如下步骤:

步骤1)对于测试集的每个特征列,求出该特征列各个的特征值以及每个特征值出现的次数;

步骤2)计算训练集样本每个特征值在测试集同一特征列所占的概率$h(a_{ij})$;

步骤3)计算训练集每个样本与测试集样本的相似度,并把相似度作为每个样本的权重;

相似度的计算方法为训练集样本每个特征值在测试集同一特征列所占的概率之和;计算公式为:

$$w_i = \sum_{j=1}^{k} h(a_{ij}) \qquad i = 1,2,\ldots,n$$

其中,n表示训练集样本的个数;

k表示特征的个数;

$w_i$表示训练集第i个样本的权重;

步骤4)基于加权的训练样本建立加权朴素贝叶斯分类器。

2.根据权利要求1所述的一种面向软件缺陷预测的加权朴素贝叶斯方法,其特征是:步骤1)具体如下:

用list存放测试集每个特征的特征值及其出现次数的元组,用HashMap来存储每个特征值及其出现次数:

$list=[dict_1,dict_2,\ldots,dict_k]$

其中,$dict_j=\{<key_1,value_1>,<key_2,value_2>,\ldots,<key_m,value_m>\}$;

k表示特征个数;

m表示测试集第j个特征中不同特征值的个数;

$dict_j$表示测试集第j个特征的特征值及其出现次数的元组;

$key_p$表示测试集某列特征的特征值;

$value_p$表示$key_p$在该特征列出现的次数。

3.根据权利要求1所述的一种面向软件缺陷预测的加权朴素贝叶斯方法,其特征是:步骤2)具体如下:

$$h(a_{ij}) = \begin{cases} value_p/count, & if\ a_{ij} = dict_j.key_p \\ 0, & else \end{cases}$$

其中,$dict_j$表示测试集第j个特征的特征值及其出现次数的元组;$key_p$表示某列特征的特征值;$value_p$表示$key_p$在该特征列出现的次数;

$dict_j.key_p$表示第j个特征中的第p个特征值;

count表示测试集样本个数;

$a_{ij}$表示训练集第i个样本的第j个特征。

4.根据权利要求1所述的一种面向软件缺陷预测的加权朴素贝叶斯方法,其特征是:步骤4)具体步骤如下:

4-1.计算先验概率;c类的加权先验概率可以重新写为:

$$P(c)=\frac{\sum_{i=1}^{n} w_i \delta(c_i,c)+1}{\sum_{i=1}^{n} w_i + n_c}$$

其中，$w_i$为训练样本i的权重；

$c_i$为训练样本i类属值；

n为训练样本总个数；

$n_c$为总类别数；在预测模型中n＝2；

δ(x,y)是指示函数；如果x＝y，则δ(x,y)＝1；若x≠y，则δ(x,y)＝0；对于类c，相同类的训练数据的样本越多，先验概率越大；

4-2.计算条件概率；根据样本加权方法，第j个特征$a_j$的条件概率为：

$$P(a_j|c)=\frac{\sum_{i=1}^{n} w_i \delta(a_{ij},a_j)\delta(c_i,c)+1}{\sum_{i=1}^{n} w_i \delta(c_i,c) + n_j}$$

其中，$a_{ij}$为第i个训练样本中的第j个特征的值；

$n_j$是第j个特征的不同值的数量；

4-3.计算测试集中样本u有无缺陷的概率，从而判断测试集样本的类别c(u)；若样本u的有缺陷概率大于无缺陷概率，则视样本u的类别为有缺陷；否则，视为无缺陷；类别c(u)公式如下：

$$c(u)=\arg max\, P(c|u)=\arg max \frac{P(c)\prod_{j=1}^{k} P(a_j|c)}{\sum_{c\in C} P(c)\prod_{j=1}^{k} P(a_j|c)}。$$

# 一种面向软件缺陷预测的加权朴素贝叶斯方法

## 技术领域

[0001] 本发明是对朴素贝叶斯的一种优化处理方法，旨在使用该技术后的朴素贝叶斯在进行软件缺陷预测时能够取得更加准确的分类预测效果，具体涉及一种面向软件缺陷预测的加权朴素贝叶斯方法。

## 背景技术

[0002] 软件缺陷预测是软件开发中非常重要的环节，能够减少软件开发中的常见缺陷，降低开发成本。软件缺陷预测也是近年来软件工程中最活跃的研究问题之一。其中分类器的性能会严重影响软件缺陷预测的准确性。

[0003] 缺陷预测最常用的分类器之一是朴素贝叶斯，尽管朴素贝叶斯比较简单，但是通常比更复杂的分类模型表现更好。然而，训练集所有样本对构建模型的贡献往往是不同的，与测试集样本相似性越高的样本构建的预测模型更加准确。朴素贝叶斯并没有考虑到训练集与测试集样本相似性对缺陷预测性能的影响，因此，本发明对朴素贝叶斯做出改进，根据训练集与测试集样本特征值的相似性，对训练集样本进行加权，高相似性的样本被赋予了更高的权重，然后在加权的训练样本上建立朴素贝叶斯模型，本发明称其为加权朴素贝叶斯方法。

## 发明内容

[0004] 本发明针对朴素贝叶斯提出了一种改进的加权朴素贝叶斯方法，该方法根据训练集样本与测试集样本的相似度，对训练集样本加权，高相似性的样本权重更大。在加权的基础上建立加权朴素贝叶斯模型。

[0005] 本发明具体包括以下步骤：

[0006] 步骤1)对于测试集的每个特征列，求出该特征列各个的特征值以及每个特征值出现的次数。

[0007] 用list存放每个特征的特征值及其出现次数的元组：

[0008] $list=[dict_1,dict_2,...,dict_k]$

[0009] 其中，$dict_j=\{<key_1,value_1>,<key_2,value_2>,...,<key_m,value_m>\}$；

[0010] k表示特征个数；

[0011] m表示测试集第j个特征中不同特征值的个数；

[0012] $dict_j$表示测试集第j个特征的特征值及其出现次数的元组；

[0013] $key_p$表示测试集某列特征的特征值；

[0014] $value_p$表示$key_p$在该特征列出现的次数。

[0015] 步骤2)计算训练集样本每个特征值在测试集同一特征列所占的概率。

[0016] $$h(a_{ij})=\begin{cases}value_p/count, & if\ \ a_{ij}=dict_j.key_p\\ 0, & else\end{cases}$$

[0017] 其中，$dict_j$表示测试集第j个特征的特征值及其出现次数的元组；$key_p$表示某列特

征的特征值；$value_p$表示$key_p$在该特征列出现的次数；

[0018] $dict_j.key_p$表示第j个特征中的第p个特征值；

[0019] count表示测试集样本个数；

[0020] $a_{ij}$表示训练集第i个样本的第j个特征。

[0021] 步骤3)计算训练集每个样本与测试集样本的相似度，并把相似度作为每个样本的权重。相似度的计算方法为训练集样本每个特征值在测试集同一特征列所占的概率之和。计算公式为：

[0022] $$w_i = \sum_{j=1}^{k} h(a_{ij}) \qquad i = 1,2,\ldots,n$$

[0023] 其中，n表示训练集样本的个数；

[0024] k表示特征的个数；

[0025] $w_i$表示训练集第i个样本的权重；

[0026] 步骤4)基于加权的训练样本建立加权朴素贝叶斯分类器。

[0027] 4-1.计算先验概率。c类的加权先验概率可以重新写为：

[0028] $$P(c) = \frac{\sum_{i=1}^{n} w_i \delta(c_i, c) + 1}{\sum_{i=1}^{n} w_i + n_c}$$

[0029] 其中，$w_i$为训练样本i的权重；

[0030] $c_i$为训练样本i类属值；

[0031] n为训练样本总个数；

[0032] $n_c$为总类别数。在缺陷预测模型中n＝2。

[0033] δ(x,y)是指示函数。如果x＝y，则δ(x,y)＝1；若x≠y，则δ(x,y)＝0。对于类c，相同类的训练数据的样本越多，先验概率越大。

[0034] 4-2.计算条件概率。根据样本加权方法，第j个特征$a_j$的条件概率为：

[0035] $$P(a_j|c) = \frac{\sum_{i=1}^{n} w_i \delta(a_{ij}, a_j) \delta(c_i, c) + 1}{\sum_{i=1}^{n} w_i \delta(c_i, c) + n_j}$$

[0036] 其中，$a_{ij}$为第i个训练样本中的第j个特征的值；

[0037] $n_j$是第j个特征的不同值的数量。

[0038] 4-3.计算测试集中样本u有无缺陷的概率，从而判断测试集样本的类别c(u)。若样本u的有缺陷概率大于无缺陷概率，则视样本u的类别为有缺陷；否则，视为无缺陷。公式如下：

[0039] $$c(u) = \arg max\, P(c|u) = \arg max \frac{P(c)\prod_{j=1}^{k} P(a_j|c)}{\sum_{c \in C} P(c)\prod_{j=1}^{k} P(a_j|c)}$$

[0040] 本发明的有益效果：

[0041] 1、该技术考虑到训练集样本与测试集样本的相似性对分类性能的影响，为高相似性的训练集样本赋予更高的权重，因此能够提高分类器的预测性能。

[0042] 2、在对训练集样本的加权过程中，考虑到了某一特征值在该特征中所占的概率大小的影响，因此权重计算方法更加准确。

## 附图说明

[0043] 图1方法流程图

## 具体实施方式

[0044] 下面根据一个简单的例子对本发明进行详细说明。本发明的整体流程图如附图1所示,具体步骤如下:

[0045] 步骤1)对于测试集的每个特征列,求出该特征列各个的特征值以及每个特征值出现的次数。

[0046] 步骤2)计算训练集样本每个特征值在测试集同一特征列所占的概率。

[0047] 步骤3)计算训练集每个样本与测试集样本的相似度,并把相似度作为每个样本的权重。

[0048] 步骤4)基于加权的训练样本建立加权朴素贝叶斯分类器。

[0049] 进一步,假设训练集有五个样本,分别为{{2,3,6,1},{1,4,5,1},{3,2,6,-1},{4,3,4,-1},{2,4,6,-1}},其中前三列为三个特征,最后一列是标签列。1表示有缺陷,-1表示无缺陷。测试集有四个样本,分别为{$u_1$={1,3,5},$u_2$={2,3,4},$u_3$={1,4,5},$u_4$={2,3,5}}。

[0050] 在步骤1中,对于测试集的每个特征列,求出该特征列各个的特征值以及每个特征值出现的次数。用list存放每个特征的特征值及其出现次数的元组:

[0051] list=[$dict_1$,$dict_2$,...,$dict_k$]

[0052] 其中,$dict_i$={<$key_1$,$value_1$>,<$key_2$,$value_2$>,...,<$key_n$,$value_m$>};

[0053] 因此,$dict_1$={<1,2>,<2,2>}

[0054] $dict_2$={<3,3>,<4,1>}

[0055] $dict_3$={<4,1>,<5,3>}

[0056] list=[$dict_1$,$dict_2$,$dict_3$]

[0057] 进一步,在步骤2中,计算训练集样本每个特征值在测试集同一特征列所占的概率。

[0058] $$h(a_{ij})=\begin{cases} value_p/count, & if\ \ a_{ij}=dict_j.key_p \\ 0, & else \end{cases}$$

[0059] 其中,count表示测试集样本个数;

[0060] $a_{ij}$表示训练集第i个样本的第j个特征

[0061] 因此,

[0062] h($a_{11}$)=0.5;h($a_{12}$)=0.75;h($a_{13}$)=0;

[0063] h($a_{21}$)=0.5;h($a_{22}$)=0.25;h($a_{23}$)=0.75;

[0064] h($a_{31}$)=0;h($a_{32}$)=0;h($a_{33}$)=0;

[0065] h($a_{41}$)=0;h($a_{42}$)=0.75;h($a_{43}$)=0.25;

[0066] h($a_{51}$)=0.5;h($a_{52}$)=0.25;h($a_{53}$)=0;

[0067] 进一步,在步骤3中,计算训练集每个样本与测试集样本的相似度,并把相似度作为每个样本的权重。相似度的计算方法为训练集样本每个特征值在测试集同一特征列所占的概率之和。计算公式为:

[0068] $$w_i = \sum_{j=1}^{k} h(a_{ij}) \qquad i = 1,2,\ldots,n$$

[0069] 其中，n表示训练集样本的个数；

[0070] $w_i$表示第i个样本的权重；

[0071] 因此，$w_1$＝0.5+0.75+0＝1.25

[0072] $w_2$＝0.5+0.25+0.75＝1.5

[0073] $w_3$＝0+0+0＝0

[0074] $w_4$＝0+0.75+0.25＝1

[0075] $w_5$＝0.5+0.25+0＝0.75

[0076] 进一步，在步骤4中，基于加权的训练样本建立加权朴素贝叶斯分类器。对于测试样本{1,3,5}，求其类标签。

[0077] 4-1.计算先验概率。c类的加权先验概率可以重新写为：

[0078] $$P(c) = \frac{\sum_{i=1}^{n} w_i \delta(c_i, c) + 1}{\sum_{i=1}^{n} w_i + n_c}$$

[0079] 其中，$w_i$为训练样本i的权重；

[0080] $c_i$为训练样本i类属值；

[0081] n为训练样本总个数；

[0082] $n_c$为总类别数。在预测模型中n＝2。

[0083] δ(x,y)是指示函数。如果x＝y，则δ(x,y)＝1；若x≠y，则δ(x,y)＝0。对于类c，相同类的训练数据的样本越多，先验概率越大。

[0084] 因此，$n_c$＝2；P(1)＝0.58，P(2)＝0.42

[0085] 4-2.计算条件概率。根据样本加权方法，第j个特征$a_j$的条件概率为：

[0086] $$P(a_j|c) = \frac{\sum_{i=1}^{n} w_i \delta(a_{ij}, a_j) \delta(c_i, c) + 1}{\sum_{i=1}^{n} w_i \delta(c_i, c) + n_j}$$

[0087] 其中，$a_{ij}$为第i个训练样本中的第j个特征的值；

[0088] $n_j$是第j个特征的不同值的数量。

[0089] 因此，$n_1$＝4；$n_2$＝3；$n_3$＝3；

[0090] P($a_1$＝1|1)＝0.37；P($a_2$＝3|1)＝0.39；P($a_3$＝5|1)＝0.43；

[0091] P($a_1$＝1|-1)＝0.17；P($a_2$＝3|-1)＝0.42；P($a_3$＝5|-1)＝0.21；

[0092] 4-3.计算测试集中样本u有无缺陷的概率，从而判断测试集样本的类别。若样本u的有缺陷概率大于无缺陷概率，则视样本u的类别为有缺陷；否则，视为无缺陷。公式如下：

[0093] $$c(u) = \arg max\, P(c|u) = \arg max \frac{P(c) \prod_{j=1}^{k} P(a_j|c)}{\sum_{c \in C} P(c) \prod_{j=1}^{k} P(a_j|c)}$$

[0094] 因此，P(1|$u_1$)＝0.93；P(-1|$u_1$)＝0.07

[0095] 因此，测试集样本$u_1$＝{1,3,5}的类标签为1。

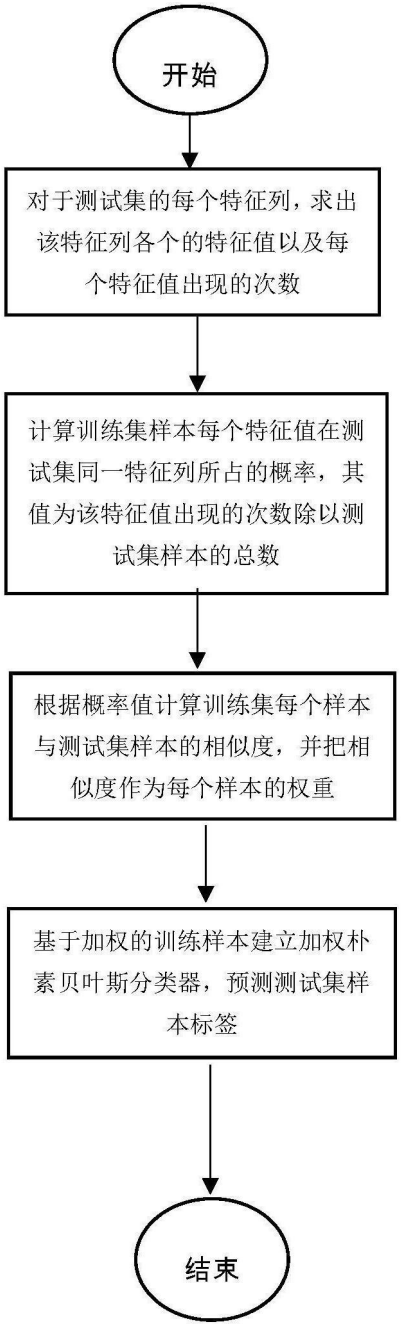

开始
对于测试集的每个特征列，求出该特征列各个的特征值以及每个特征值出现的次数
计算训练集样本每个特征值在测试集同一特征列所占的概率，其值为该特征值出现的次数除以测试集样本的总数
根据概率值计算训练集每个样本与测试集样本的相似度，并把相似度作为每个样本的权重
基于加权的训练样本建立加权朴素贝叶斯分类器，预测测试集样本标签
结束

图1