基于信息增益特征关联树的文本特征选择算法

任永功 杨 雪 杨荣杰 胡志冬

(辽宁师范大学计算机与信息技术学院 大连 116029)

摘 要 传统的信息增益算法在类和特征项分布不均时,分类性能明显下降。针对此不足,提出了一种基于信息增益特征关联树的文本特征选择算法(UDsIG)。首先,对数据集按类进行特征选择,降低类分布不均时对特征选择的影响。其次,利用特征分布均匀度改善特征项在类内分布不均对特征选择的干扰,并采用特征关联树模型对类内特征进行处理,保留强相关特征,删除弱相关和不相关特征,降低特征冗余度。最后,使用类间加权离散度的信息增益公式进一步计算,得到更优特征子集。通过对比实验表明,选取的特征具有更好的分类性能。

关键词 特征选择,特征关联树,信息增益值,不平衡数据集,离散度

中图法分类号 TP301.6 **文献标识码** A

Text Feature Selection Methods Based on Information Gain and Feature Relation Tree

REN Yong-gong YANG Xue YANG Rong-jie HU Zhi-dong

(School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)

Abstract Due to the maldistribution of classes and features, the classification performance of traditional information gain algorithm will decline sharply. Considering that, a text feature selection method UDsIG was proposed which is based on the information gain. Firstly, because the feature selection may be influenced when the classes is unevenly distributed, we selected features based on class. Secondly, we used feature distribution uniformity to improve the influence on feature selection process when features are uneven distributed in the class. Then we adopt the feature relation tree model to deal with the class features, retain strong correlation features and delete the weak correlation and irrelevant ones. At last, we got the best feature subset by using of information gain formula which is based on weighted dispersion. The comparison experiment shows that the method has better classification performance.

Keywords Feature selection, Feature relation tree, Information gain, Imbalanced dataset, Dispersion

1 引言

随着网络普及率的提高,我国网络用户数量突增,各式各样的网站中蕴涵着海量的中文信息,然而这些信息绝大多数是以文本的形式存在的,这对信息处理提出了新的挑战。由于中西方语言存在着巨大差异,国外在文本分类方面的研究成果无法直接应用到中文文本分类中,因此对中文文本分类技术的研究具有非常重要的现实意义。在文本分类系统中,分类算法通常采用向量空间模型(Vector Space Model, VSM)[1]来表示文本,然而向量空间的高维性和文档表示向量的稀疏性限制了文本分类的处理速度。因此,特征选择显得尤为重要。特征选择通过降低特征空间维度以及去除噪音特征等方法来提高分类效率及精度。目前文本分类算法中常用的特征选择方法有交互信息(MutualInformation,简称MI)、信息增益(Information Gain,简称IG)、X2统计量(Chi-square,简称CHI)、特征权(Term Strength,简称TS)、期望交叉熵(ExpectedCrossEntropy,简称ECE)、文本证据权(WeightofEvi-dence,简称WE)等。

IG算法是一种有效的特征选择算法,它伴随着文本分类的出现而出现,后又被广泛应用于其他领域,如图像处理、生物信息学等。通过现有信息增益算法的研究表明,IG算法能有效地选出关键特征、剔除无关特征[10]。但是现有的IG算法仍然存在不足,其一:未对特征项的词频信息给予足够的重视;其二:对类和特征项分布不均的情况考虑不充分。因此为提高IG算法的性能,本文做了如下改进,首先针对类内特征,利用特征分布均匀度来降低特征分布不均的影响,通过特征关联树删除冗余特征,对类内特征做进一步筛选;其次针对特征项在类间分布不均的情况,利用类间加权离散度来选择更精确的特征子集,使分类效率和精度得到明显提高。

2 相关定义

定义1(强相关) 设F是一个特征子集,f_i是一个特征,$S_i=F-\{f_i\}$中特征f_i是强相关的当且仅当

$$P(C\backslash f_i,S_i)\neq P(C\backslash S_i) \tag{1}$$

强相关特征是对类的分布构成影响的特征,如果缺少了必然改变类的分布情况,则其是最优子集的一部分。

到稿日期:2012-12-23 返修日期:2013-05-05 本文受辽宁省计划项目(2012232001),辽宁省自然科学基金(201202119)资助。

任永功(1972—),男,博士,教授,主要研究方向为数据挖掘、图像处理技术等,E-mail:renyg@dl.cn;杨 雪(1987—),女,硕士生,主要研究方向为数据挖掘;杨荣杰(1983—),女,硕士生,主要研究方向为数据挖掘;胡志冬(1984—),男,硕士生,主要研究方向为数据挖掘。

定义 2(弱相关) 特征 f_i 是弱相关的当且仅当

$P(C\backslash f_i,S_i)=P(C\backslash S_i)$且$\exists S_i'\subset S_i$,则 $P(C\backslash f_i,S_i')\neq P(C\backslash S_i')$ (2)

弱相关特征在一定条件下影响类的分布。

定义 3(不相关特征) 特征 f_i 是无相关的当且仅当

$\forall S_i'\subseteq S_i, P(C|f_i,S_i')=P(C|S_i')$ (3)

不相关特征不影响类的分布情况,因此可以删掉。

定义 4(相对冗余) 设 F 是一个特征集合,集合 M 是 F 中一个子特征集合,如果特征 f_i 在 $M-\{f_i\}$ 中存在

$P(F-M_i-\{f_i\}|f_i,M_i)=P(F-M_i-\{f_i\}|M_i)$ (4)

那么 f_i 相对于 M 是冗余的。

3 基于信息增益的改进算法 UD$_s$IG

3.1 信息增益简介

信息增益(Information Gain,IG)是通过计算某一特征能为分类带来多少有用的信息来衡量特征对于分类的重要程度,以某特征在文本中出现前后的信息熵之差作为权值,进而筛选出最优的分类特征子集。特征的信息增益越大,其包含的信息量也就越大,在分类中越重要。信息增益的公式如下:

$$IG(W)=-\sum_t P(C_t)\log P(C_t)+P(W)\sum_t P(C_t/W)\log P(C_t/W)+P(\overline{W})\sum_t P(C_t/\overline{W})\log P(C_t/\overline{W}) \quad (5)$$

式中,t 表示文档类别总数,$P(C_t)$表示 C_t 类文档在语料库中出现的概率;$P(W)$表示特征 W 在文本中出现的概率;$P(C_t/W)$表示文档包含特征 W 时属于 C_t 类的概率,$P(\overline{W})$表示特征 W 在文本中不出现时的概率,$P(C_t/\overline{W})$表示文本不包含特征 W 时属于 C_t 类的概率。

3.2 信息增益的不足

首先,信息增益方法是从整个训练集角度进行特征赋权的,计算的是特征在诸类别上出现与否的后验概率,从整个训练集来看这种特征选择模式对于文本向量表示是合理的,但是在选择各类的类别特征集合时,不能体现出该类别相应的类别特征所特有的信息,所以不适合构造类别特征向量。其次,传统信息增益特征提取方法更多地关注了文档数,而对词频没有给予足够的重视。此外,特征间相关性也没有被考虑,这导致特征预测能力被削弱,无法选出最有效的特征。最后,由于考虑了特征项不出现的情况,当类别分布不均衡或者特征项分布不均衡时,绝大多数特征项在某些类别中不出现,此时的信息增益值主要由公式的后半部分决定,因此会使得在一个类别中出现次数不多而在其他类别中频繁出现的特征被选出来,而不倾向于选取在一个类别中出现较多而在其他类别中出现较少的更具代表性的特征项。从而导致信息增益效果大大下降。

现实中不平衡数据集的现象普遍存在,为了使 IG 算法能够得到更广泛的应用,需要对 IG 算法现有的不足做出改进。针对 IG 算法只适用于全局变量且没有考虑词频对于特征选择的影响这一缺陷,文献[4,6]分别采用了不同的方法对 IG 算法做出改进,提出了适用于不平衡数据集的新算法。文献[4]中将 Information gain(IG)、Chi-square(CHI)、Correlation coefficient(CC)以及 Odds ratio(OR) 4 种特征选择算法进行组合,组合后的方法对不平衡数据集的特征选择具有一定的效果,但是由于它结合了多个特征选择方法,因此限制条件较多,适用性差;文献[6]中当类分布不均时,此算法精度会下降,因此适用范围小,没有普遍性。

基于传统 IG 算法以及文献[4,6]中的不足,本文分别考虑了类和特征项分布不均时的情况,对 IG 算法进行改进,提出了一种改进的基于信息增益特征关联树的文本特征选择方法。一方面,对数据集进行分类特征选择,利用特征关联树模型,降低特征在类内分布不均对特征选择的影响。另一方面,为了消除特征项在类间分布不均衡或者类分布不均衡的情况,利用加权离散度作为平衡因子,改进信息增益公式,降低传统方法因考虑特征不出现时所带来的干扰,以获取更为精确的特征子集。

3.3 基于类内特征关联树模型去除冗余特征

3.3.1 改善类内特征分布不均的影响

传统的信息增益算法只考虑了特征的文档数,对于词频没有给予足够的重视,且传统方法考察的是特征对于全局的贡献,而不能具体到某一个类中,认为在同一类文本中,出现次数越大的特征项越有代表性,这是基于特征项在类内文本中分布均匀的假设,没有考虑到特征项在该类文本中分布均匀的程度。如果特征项在类内分布不均匀,那么它就不能很好的代表该类。例如,在类别 C_1,C_2,C_3 3 个类中,每个类都含有 5 个样本,有两个特征项 w_1,w_2。它们在类 C_1 中出现的词频总数相同,但分布的均匀程度不同,如表 1 所列。

表 1 特征文档分布情况

特征项	C_1					C_2					C_3				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
w_1	3	3	3	3	3	0	0	0	0	0	0	0	0	0	0
w_2	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0

由传统的信息增益式(5)得:$IG(w_1)=0.730$,$IG(w_2)=0.766$。从结果可知特征分布是否均匀严重影响特征的提取,w_1 在类中分布得更均匀,相对于 w_2 而言应当具有更强的类区分性。但若根据 IG 值的大小,则 w_1 更容易被过滤掉。

为了解决这一问题,本文使用特征分布均匀度对类别特征集进行初步筛选,其公式如下所示:

$$U(t)=\left(1-\frac{\sqrt{\frac{1}{D_i-1}\sum_{j=1}^{D_i}(tf_j(t)-\overline{tf_j(t)})^2}}{\overline{tf_j(t)}}\right) \quad (6)$$

式中,设类 C_t 的总文档数为 D_i,$\overline{tf_j(t)}$代表特征 t 在类 C_t 文档中的平均词频,$tf_j(t)$代表特征项 t 在第 j 篇文档中出现的频度。首先统计类别 C_t 中特征出现的频数,选取频数大于 $D_i/3$ 的特征,然后利用式(6)进行筛选,并按照 $U(t)$值降序排列,选取前 800 个特征组成分类特征子集 W。

3.3.2 建造类内特征关联树模型去除特征冗余

目前大多数的特征选择方法是基于特征的频数或特征与类别间的相关性进行的,而特征项之间的相关性信息却很少被作为特征选择的依据。建造模型的目的是为了构造"特征-特征"关联矩阵[7]。这种特征矩阵在特征空间中表示为任意两个节点之间都有连线。在同一类别中,我们需要的是关联度较大的特征,对于相关性低,且在统计意义上很少共现的,则没有存在的意义。因此,必须通过剪枝,删除关联度小的路径,仅保留关联值大的那些路径,降低特征间的冗余。我们把这个模型称为"特征关联树"模型[9]。

"特征关联树"模型是将特征向量按照特征性质划分为

N 个小的特征空间，从空间中任意选取一个特征作为树根，按照该特征的性质搜索与该特征相关联的特征作为孩子，计算特征间的关联相似度，并按照相似度的降序排列选择前 K 个关联特征，组成一个最优特征子集，对关联特征树进行剪枝。同理再从 $N-1$ 个特征空间中选取一个特征，搜索与该特征相关联的特征，并进行剪枝，直到 $N=0$。此时生成的关联树的所有特征均是较优特征且相互间关联度很大。

设 $\mathrm{sim}(i,j)$ 表示两个特征 f_i 和 f_j 之间的相似度。对任意一个给定的特征 f_i，采用树状的模型来表示特征词 i 与 j 之间的关系，如图 1 中左边部分所示。图 1 中，在特征子集 F 中选取 $U(t)$ 值最大的特征 f_k 作为树根，其余的特征是树叶，两者之间树枝(即路径)的权值是两个特征之间的相似度。f_k 到 f_i 的所有路径，按照相似度大小排序：

$$\mathrm{sim}(f_k,f_1)\geqslant\cdots\geqslant\mathrm{sim}(f_k,f_i)\cdots \tag{7}$$

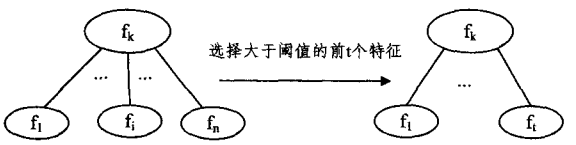

图 1　特征关联树

相似度 $\mathrm{sim}(f_k,f_i)$ 的计算方法有很多种，本文采用基于互信息的方法。

设 $F(f_1,f_2,\cdots,f_n)$ 表示初始特征向量，向量 f_i 表示第 i 个特征。图 1 中表示的是一个特征关联树元素。事实上，整个特征关联树是由很多个特征关联树元素在不同层次上搭建而成的。传统方法是从左向右依次提取出 K 个树叶(即相似度最大的前 K 个特征)，但实际中，可能存在相似度小的特征被选中加入树中的情况，这样会导致噪声的引入，因此本文提出根据设定的相似度阈值 θ、树的扩展层数 L 控制树叶个数和扩展层数，通过阈值的限制，能更好地控制关联树的噪声，缩减特征维数的规模，降低计算复杂性。

3.4　基于类间加权离散度的信息增益公式改进

传统的信息增益算法考虑了特征项不出现时对类别判定的贡献程度，但在类别分布不均衡或者特征项在类间分布不均衡的情况下，由于绝大多数特征项在某些类别中是不出现的，此时信息增益值的大小主要依赖于特征项不出现情况下所带来的信息量的大小。实验表明在数据集不平衡时，对分类的干扰过大。例如有 C_1，C_2，C_3，C_4 4 类，前 3 类都含有 5 个样本，C_4 含有 1 个样本。有 3 个特征项 w_1，w_2，w_3，分布如表 2 所列。

表 2　特征文档分布情况

特征项	C_1					C_2					C_3					C_4
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1
w_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
w_2	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0
w_3	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2

由式(6)得：$IG(w_1)=0.337$，$IG(w_2)=0.896$，$IG(w_3)=0.393$。从结果可知数据集平衡程度严重影响特征的提取，w_1 相对于 w_3 具有更强的类区分性，但根据信息增益值筛选特征，w_1 更容易被过滤掉。

离散度是用来观测变量各个取值之间的差异程度的重要指标，以衡量一组数据的波动大小。这里，我们考虑类别比重，用以平衡类别分布不均衡造成的影响。设有 m 个类，各个类别所占的比重分别为 $p_1,p_2,\cdots,p_n$，$\sum p_i=1$。改进后的加权离散度式如式(8)所示：

$$D_s=\frac{\sqrt{\frac{p_m}{1-p_m}\sum_{i=1}^{m}p_i(tf_i(t)-\overline{tf(t)})^2}}{tf(t)} \tag{8}$$

式中，$tf_i(t)$ 表示特征 t 在第 i 类中出现的频度，m 为类别总数，$\overline{tf(t)}$ 为 t 在各类中出现频度的平均值，$tf(t)$ 表示 t 在所有类别中出现的总频度，P_m 表示含有文档数最多的类别的比重。$D_s=0$ 的充分必要条件是特征在各类别中分布均衡，即 $tf_1(t)=tf_2(t)=\cdots=tf_n(t)$；$D_s=1$ 的充分必要条件是特征在类间分布不均衡。特征选择最终筛选的是类中出现概率大且其它类别中不出现的特征。故 D_s 值越大，表明特征仅在一个类别中出现的可能性越大，分类能力越强。

我们将离散度作为平衡因子加入到信息增益公式中，用以平衡由于特征项在类间分布不均衡或者类分布不均衡时所带来的分类干扰。改进后的信息增益公式如下：

$$D_sIG(t)=-\sum_i P(C_i)\log P(C_i)+P(t)\sum_i P(C_i|t)\log P(C_i|t)+P(\bar{t})\sum_i P(C_i|\bar{t})\log P(C_i|\bar{t})\times D_s \tag{9}$$

由式(9)得，$IG(w_1)=1.46$，$IG(w_2)=1.53$，$IG(w_3)=1.39$。其中 $IG(w_1)>IG(w_3)$ 且 $IG(w_2)>IG(w_3)$，结果更符合实际情况。离散度的引入，修正了传统方法中数据集不平衡所带来的干扰。

3.5　基于信息增益特征关联树的特征选择算法的描述

基于信息增益特征关联树的特征选择算法描述如下：

算法 1　基于类内特征关联树模型训练过程描述如下：

输入：类内文本 C_{text}，阈值 θ，扩展层数 L

输出：特征关联树 T

1. 预处理 C_{text}，选取特征出现频数 S> $D_t/3$ 的特征 M[m]，并初始化
2. 利用式(6)计算 $M_i(i=1,2,\cdots,m)$ 的 U(t)值并降序排列；
3. 选取前 800 个特征组成分类特征子集 W；
4. Set SubTree(W_i)= ϕ；/ * 读入 W 中 U(t)值最大的特征 W_i 为根节点，构建最优子树 SubTree(W_i) * /
5. Initialize G=NULL；
6. For(int j=0,j<i,j++)；
7. While(L<=5)
8. 计算特征间相似度值，选取 $\mathrm{Sim}(W_i,F_j)>0.6$ 的节点存入 S 向量空间中，并将第一个相似度最大的特征加入 G 中
9. 　　F_j=getnext(S)；
10. 　　G←F_j，F_j=children(SubTree(T))；
11. Goto step4；
12. L=L++；
13. End for；
14. Return G，Subtree(T)；

算法 2　利用改进的信息增益公式选取最优特征子集

输入：所有类别建立特征关联树所得特征 S

输出：S_{last} 特征子集

1. S′=Delete(r)；/ * r 为类别特征 S 中的重复特征 * /
2. 根据式(9)计算特征子集 S′中特征的信息增益值 $D_sIG(S_i')$；
3. 将 S′中 $D_sIG(S_i')>0.8$ 的特征存入 S_{last} 中；
4. 输出 S_{last} 特征子集；

为降低不平衡数据集对特征选择的影响，首先使用平均度和特征关联树分类选择特征，然后通过基于加权离散度改进的信息增益公式对特征进一步筛选。对于不同的训练数据，其计算出的信息增益值不同，则设置的阈值也不同。阈值

太小，不能起到筛选特征的作用，太大则造成过分筛选，删除对文本分类重要的特征。本文经过多次实验对比，在阈值 θ 设置为 0.8 时得到最佳实验结果。最后使用选取的特征进行分类实验，以验证本文改进的特征选择算法的有效性。

4 实验结果及分析

4.1 评价指标

文本分类的评测指标采用文本分类评测标准中的平均准确率、平均召回率和 F1 值。各评价参数定义如下：

1)平均准确率

分类的准确率=分类正确文本/分类的实际文本数

$$MacroP=\frac{1}{n}\sum_{j=1}^{n}P_j \tag{10}$$

式中，n 为总的分类数，P_j 为第 j 类的准确率。

2)平均召回率

分类的召回率=分类正确文本/分类应有的文本数

$$MacroR=\frac{1}{n}\sum_{j=1}^{n}R_j \tag{11}$$

式中，n 为总的分类数，R_j 为第 j 类的召回率。

3)平均 F1 值

$$MacroF1=\frac{MacroP\times MacroR\times 2}{MacroP+MacroR} \tag{12}$$

4.2 试验分析

本文对上述特征选择方法的分类效率进行了实验，实验数据选取了复旦大学中文语料库中的 2655 篇文本，包括历史 800 篇、文学 55 篇、艺术 800 篇、教育 100 篇、计算机 900 篇，其中 70%用来训练，30%用来测试。首先使用 ictclas50 分词包对所选文本进行分词处理，之后提取文本中的名词、动词、形容词和量词，去除其中的停用词和无用词，使用本文提出的特征选择算法选取特征，最后使用 KNN 文本分类器对测试样本进行分类。

针对不平衡数据集对分类效果的影响，首先做了如下实验。在复旦大学的语料库中，文学类文本一共 67 篇，计算机类文本 2715 篇，这两个类构成一个不平衡数据集，在文学类中随机选取 50 篇文档，计算机类文档按以下比例随机抽取，如表 3 所列。

表 3 不平衡数据集的实验数据

非平衡比(文学：计算机)	1：1	1：5	1：10	1：20
文档数(计算机)	50	250	500	1000

由于该实验只有两个类别，因此 KNN 算法的 K 值取 1，选取 1000 个特征，采取代表查全率和查准率调和均值的 F1 值作为评价指标，下面是不同非平衡比下传统的 IG 算法和本文提出 UDsIG 算法的实验结果对比，如表 4 所列。

表 4 IG 和 UDsIG 特征选择方法在不平衡数据集上的 F1 值对比(%)

非平衡比	1：1		1：5		1：10		1：20	
特征选择	文学	计算机	文学	计算机	文学	计算机	文学	计算机
IG	94.11	95.91	82.35 96.78	83.33	98.90	76.47	99.19	
UD_sIG	94.11	95.91	92.15	98.79	93.87	99.20	95.83 99.60	

由实验数据可知，在数据集不平衡时，传统 IG 算法下，文学类的 F1 值明显下降，而 UD_sIG 方法中，文学类文本的 F1 值相对比较稳定。

因此，在整个训练集上进行实验，对每类出现频数大于 1/3 文档数的特征分别进行传统 IG 算法、特征关联树模型与传统 IG 算法相结合的 U(t)-IG 以及改进的 UD_sIG 3 种算法的计算，然后在每类中选取权值高的前 200 个特征，使用 KNN[8] 分类器对测试集进行分类，通过对比 3 种情况下的分类精度验证了本文特征提取算法能有效提高文本分类精度。其中，IG 特征选择算法与 U(t)-IG 特征选择算法以及 UD_sIG 算法的性能比较如图 2 所示。

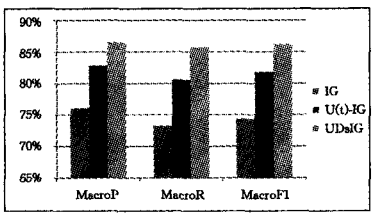

图 2 3 种特征选择算法的性能比较

由图 2 可知，U(t)-IG 特征选择算法相对于传统的 IG 特征选择算法在 MacroP、MacroR、MacroF1 3 个方面都有很大的提高。这是由于 U(t)-IG 特征选择算法较传统的 IG 算法而言，首先分类进行特征选择，在此基础上，采用类内特征分布均匀度降低特征项在类内分布不均衡时对分类带来的影响，然后根据特征相关性大小，建立特征关联树模型，删除冗余特征，降低特征维度。UD_sIG 特征选择算法相对于 IG 特征选择算法、U(t)-IG 特征选择算法又有一定提高。这是因为 UD_sIG 特征选择算法在 U(t)-IG 算法的基础上考虑了特征在类间分布的离散程度，对信息增益公式进行了改进，利用加权的离散度公式，将类别比重、词频信息考虑在其中，降低了特征项在类间分布不均衡时或者类分布不均衡所带来的干扰，使提取的特征对区分类别具有更高的贡献率，以得到更为理想的分类效果。

结束语 本文在研究传统信息增益算法和改进算法的基础上，对其存在的不足进行了优化。通过分类选择特征，改善数据集不平衡对特征选择的影响。在类内，利用特征关联树模型去除冗余特征，降低特征维度；在类间，利用类别分布权重信息对离散度公式进行加权，利用加权离散度改进信息增益公式，减少因类分布不均对分类造成的影响。实验结果表明，在中文文本分类过程中，类和特征分布不均时，改进后的信息增益算法的分类性能有很大的提高。

参 考 文 献

[1] Kao C C. Design of echo cancellation and noise elimination for speech enhancement[J]. IEEE Transactions on Consumer Electronics, 2003, 49

[2] Ng H, Goh W, Low K. Feature selection, perceptron learning and a usability case study for text categorization [C]// Proceedings of the 20th ACM International Confer-ence onResearch and Development in InformationRetrieval(SIGIR-97). 1997: 67-73

[3] Xu Yan, Chen Lin. Term-frequency Based Feature Selection Methods for Text Categorization[C]// Proceedings of the 2010 Fourth International Conference on Genetic and Evolutionary Computing. Dec. 2010

[4] J Xian, L Pei-yu, G Wei, et al. An algorithm application in intrusion forensics based on improved information gain [C] // 3rd Symposium on Web Society(SWS)2011. 2011

[5] Wang Zi-qiang, Zhang De-xian. Feature Selection in Text Classification Via SVM and LSI[J]. Lecture Notes in Computer Science, 2006, 3971: 1381-1386

[6] Yang Yu-zhen, Liu Pei-yu, Zhu Zhen-fang, et al. The Research of an Improved Information Gain Method Using Distribution Information of Terms[C]// IEEE International Symposium. 2009: 938-941

[7] 崔自峰,徐宝文,张卫峰. 一种近似 Markov Blanket 最优特征选择算法[J]. 计算机学报,2007,30(12):2074-2081

[8] Hu Qing-hua, Yu Da-ren, Xie Zong-xia. Neighborhood classifiers [J]. Expert Systems with Applications, 2008, 34(2): 866-876

[9] 刘海峰,王元元,姚泽清. 文本分类中一种基于选择的二次特征降维方法[J]. 情报学报,2009,28(1):23-27

[10] 徐燕,李锦涛,王斌,等. 基于区分类别能力的高性能特征选择方法 [J]. 软件学报,2008,19(1):82-89

[11] 周城,葛斌,唐九阳,等. 基于相关性和冗余度的联合特征选择方法[J]. 计算机科学,2012,39(4):181-184

[12] 刘庆和,梁正友. 一种基于信息增益的特征优化选择方法[J]. 计算机工程与应用,2011,47(12):130-136

(上接第 220 页)

解:①分别求出两个可逆逻辑函数的二维辅因子码值向量,分别为:$V_w(F)=[(3,3,3),(3,2,2),(3,3,2)]$,$V_w(G)=[(3,3,2),(3,2,2),(3,3,3)]$,排序后有:$V_w(F)=V_w(G)=[(3,3,3),(3,3,2),(3,2,2)]$。

②求出各个输出分量的 RM 展开式有:F 的输出分量为 $f_3=x_1x_2\oplus x_1x_3\oplus x_2x_3$,$f_2=x_1\oplus x_2\bar{x}_3$,$f_1=x_1\oplus x_1x_3\oplus\bar{x}_2x_3$;$G$ 的输出分量为 $g_3=\bar{x}_1\oplus\bar{x}_1x_2\oplus x_2\bar{x}_3=1\oplus x_1\oplus x_1x_2\oplus x_2x_3=\bar{x}_3\oplus\bar{x}_2\bar{x}_3\oplus\bar{x}_1\bar{x}_2=1\oplus x_3\oplus\bar{x}_2x_3\oplus x_1\bar{x}_2$,$g_2=\bar{x}_1\oplus\bar{x}_2\bar{x}_3=1\oplus x_1\oplus\bar{x}_2\bar{x}_3$,$g_1=x_1x_2\oplus x_1\bar{x}_3\oplus x_2\bar{x}_3=1\oplus\bar{x}_1\bar{x}_2\oplus\bar{x}_1x_3\oplus\bar{x}_2x_3$。

③根据可逆逻辑函数输出分量的码值向量,建立输出分量之间的对应关系仅有:$\varphi:(f_3,f_2,f_1)\leftrightarrow(g_1,g_2,g_3)$。

④求对应输出分量 NP-N 等价时,所有变量映射的集合。当 f_3 和 g_1 NP-N 等价时,变量映射的集合 S_1 共有 12 个元素,分别是 (x_1,x_2,x_3) 与 $(x_1,x_2,\bar{x}_3)$ 的所有置换的对应和 (x_1,x_2,x_3) 与 $(\bar{x}_1,\bar{x}_2,x_3)$ 所有置换的对应,当 f_2 与 g_2 NP-N 等价时,变量映射的集合 S_2 为 (x_1,x_2,x_3) 分别与 $(\bar{x}_1,\bar{x}_2,x_3)$,$(\bar{x}_1,\bar{x}_3,x_2)$,$(x_1,\bar{x}_2,x_3)$,$(x_1,\bar{x}_3,x_2)$ 的对应,当 f_1 与 g_3 NP-N 等价时,变量映射的集合 S_3 为 (x_1,x_2,x_3) 分别与 $(\bar{x}_1,x_3,x_2)$,$(x_1,\bar{x}_3,x_2)$,$(x_3,x_1,\bar{x}_2)$,$(x_3,\bar{x}_1,\bar{x}_2)$ 的对应。

⑤于是,可求得:$S_1\cap S_2\cap S_3=\{(x_1,x_2,x_3)\leftrightarrow(x_1,\bar{x}_3,x_2)\}\neq\varnothing$,所以,可逆逻辑函数 F 和 G 是 NP-NP 等价的。

结束语 在可逆逻辑函数的综合中,分类可以使模块重复使用。因此,对可逆逻辑函数分类的研究是必要的,而判断两个可逆逻辑函数是否属于同一类又是研究可逆逻辑函数分类的重要部分。先计算出可逆逻辑函数辅因子码值向量,并把码值向量排序后是否相同作为可逆逻辑函数是否 NP-NP 等价的初步判定,当排序后的辅因子的码值向量相同时,再建立各个输出分量之间的对应关系,然后找出各个输出分量 NP-N 等价时的变量映射集合,通过判断这些集合的交集是否为空来判断给定的可逆逻辑函数是否 NP-NP 等价。我们把 3 阶可逆逻辑函数作为研究对象,并成功解决了 3 阶可逆逻辑函数 NP-NP 等价的判定问题。

参 考 文 献

[1] Landauer R. Irreversibility and Heat Generation in the Computing Process [J]. IBM Journal of Research and Development, 1961(5): 183-191

[2] Bennett C H. Logical reversibility of computation [J]. IBM Journal of Research and Development, 1973, 17(6): 525-532

[3] Maslov D, Dueck G W, Miller D M. Synthesis of Fredkin-Toffoli Reversible Networks [J]. IEEE Transactions on Very Large Scale Integration(VLSI)Systems, 2005, 13(6): 765-769

[4] Fazel K, Thornton M, Rice J E. ESOP-based Toffoli Gate Cascade Generation[C]// IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. Aug. 2007: 206-209

[5] Li W Q, Chen H W, Li Z Q. Application of semi-template in reversible logic circuit [C]// Proceedings of the 11th International Conference on CSCWD. Melbourne, Australia, 2007: 155-161

[6] Song Xiao-yu, Yang Guo-wu, Perkowski M, et al. Algebraic Characterization of Reversible Logic Gates [J]. Theory of Computing Systems, 2006, 39(2): 311-319

[7] Shende V V, Prasad A K, Markov I L, et al. Synthesis of reversible logic circuits [J]. IEEE Trans on Circuits and Systems I, 2003, 22(6): 723-729

[8] Yang G W, Song X Y, Perkowski M, et al. Fast synthesis of exact minimal reversible circuits using group theory [J]. Proceedings of IEEE ASP-DAC, 2005(2): 18-21

[9] Rice J E. Considerations for Determining a Classification Scheme for Reversible Boolean Function [R]. TR-CSJR2-2007

[10] Tsai C C, Marek-Sadowska M. Boolean Functions Classification via Fixed Polarity Reed-Muller Forms [J]. IEEE Trans. Computers, 1997, 46(2): 173-186

[11] Mozammel H A, Khan A. Quantum Logic Circuit For Generating Fixed-Polarity Reed-Muller Coefficients [C]// 4th International Conference on Electrical and Computer Engineering. December 2006: 141-144

[12] Hirayama T, Takahashi M, Nishitani Y. Simplification of Exclusive-or Sum-of-Products Expressions Through Function Transformation [C]// Circuits and Systems, IEEE Asia Pacific Conference. Dec. 2006: 1480-1483

[13] 刘永才,张卫. 布尔方法论 [M]. 上海:上海科学技术文献出版社,1993

[14] Perkowski M, Joziwak L, Mixhchenko A, et al. A General Decomposition for Reversible Logic[C]// Proceedings of the International Workshop on Methods and Representations(RM). August 2001: 119-138