不平衡数据集下特征词两面性的新型降维算法*

付 鑫，王洪国†，邵增珍，杜秋霞
（山东师范大学 信息科学与工程学院 山东省物流优化与预测工程技术研究中心，济南 250014）

摘 要：传统 DFS 特征选择算法在降维处理时既未考虑样本分布不均的情况，又未涉及负特征词对类别的影响。综合考虑 DFS 的缺陷并进行优化处理，将 DFS 与卡方检测算法 CHI 结合，提出一种改进型特征选择算法 DFS-sCHI。引入负特证词作为类别划分的影响因子之一，解决不平衡数据集下所提特征词类别分布不均的问题。经实验分析，不平衡数据集下，DFS-sCHI 相比较于 DFS 在分类精度上有明显提高。

关键词：不平衡数据集；文本分类；特征选择；DFS-sCHI
中图分类号：TP391.1 **文献标志码**：A **文章编号**：1001-3695(2018)07-1947-03
doi:10.3969/j.issn.1001-3695.2018.07.005

Novel feature selection method based on two-side considering imbalance problem

Fu Xin, Wang Hongguo†, Shao Zengzhen, Du Qiuxia
(*Shandong Provincial Logistics Optimization & Predictive Engineering Technology Research Center, School of Information Science & Engineering, Shandong Normal University, Jinan* 250014, *China*)

Abstract: DFS(distinguishing feature selector) neither considered the situation of uneven distribution of samples nor involved the impact of negative words for category. This paper considered the defects of DFS and optimized these defects, combined DFS with CHI to proposed DFS-sCHI that was a improve feature selection method. This method introduced the power of negative words in order to solve the problem that the feature of category was uneven selection. The experimental results indicate that the proposed method improves the classification accuracy obviously.

Key words: imbalance data set; text classification; feature selection; DFS-sCHI

由于计算机技术的飞速发展，网络中储存的无结构化文本数量越来越多。文本分类作为一种组织和处理大量无结构化文本的自动分类技术，得到广泛关注。用于分类的文本一般采用向量空间模型[1]的表示方式，但是若将样本集中的特征词都放入模型中，列向量将达到万维，此向量数据量庞大，不利于分类。解决方法之一是进行降维处理，常用的降维算法分为特征提取降维和特征选择降维两类。考虑到处理时间与难度等因素限制，本文采用特征选择算法降维。特征选择算法包含基于过滤的、基于封装的以及基于嵌入式的特征选择算法。基于过滤的算法具备最低的处理时间，受到研究者的喜爱，常见的有文档频率[2]、信息增益[3]、卡方统计量[4]、互信息[5]、改进的基尼指数[6]、泊松分布[7]等。

评价特征词重要性的降维算法多数都是基于均衡数据集，即各个类别的训练与测试样本数相同，而未考虑在不平衡数据集下的影响。所谓不平衡数据集，是指同一样本集中某些类别的样本数远大于或远小于其他类别。考虑一些实际因素，在医疗诊断、垃圾邮件分类、诈骗检测等领域，收集的样本集若各个类别的数量相同则需要耗费大量时间精力，然而不平衡数据集在各个领域广泛存在并且容易获得。因此最近几年，不平衡数据集在国内外成为研究热点[8]。

针对不平衡数据集下传统特征选择算法存在的不足，前人提出许多改进算法。闫健卓等人[9]根据卡方统计量 CHI 未考虑特征词在类间、类内的不同分布对分类的影响，引入信息熵函数，从而提高卡方检测在不平衡数据集下的分类精度；任永功等人[10]根据信息增益在平衡数据集下精度降低，提出特征关联树的概念，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；尤鸣宇等人[11]根据 IG 算法在不平衡数据集上存在的问题，提出 Im-IG 算法，通过提高小类分布在信息熵计算中的权重，优先选入有利于小类正确分离的特征；徐燕等人[12]提出一种新型特征选择算法 DFICF；张玉芳等人[13]考虑特征在正类和负类中的分布性质，提出 IPR 算法。相同的策略对于不同的特征选择算法收效甚微，DFS(distinguishing feature selector)[14]作为一种新型特征选择算法，在不平衡数据集下的分类效果并未引起广泛关注。本文通过引入特征词正负相关性的概念，从特征词的两面性考虑，筛选合适的特征词数量分布于各个类别，克服 DFS 类别分布不均导致分类下降的问题。

1 不平衡数据集下新型降维算法

1.1 DFS 算法困境

DFS 是一种全局特征选择算法，通过计算特征词在各个类别中的重要性程度之和，得到此特征词总的评价指标。某个特征词 t 重要性程度的计算公式如下所示：

$$\mathrm{DFS}(t)=\sum_{i=1}^{M}\frac{P(C_i/t)}{P(\bar{t}/C_i)+P(t/\bar{C}_i)+1} \tag{1}$$

其中：C_i 表示文本类别；$P(C_i/t)$ 表示特征词 t 在 C_i 中出现的比例；$P(\bar{t}/C_i)$ 表示 C_i 类别下特征词 t 出现的比例；$P(t/\bar{C}_i)$ 表示 C_i 类别下特征词 t 出现的比例。

在不平衡数据集下，样本稀有类别中即使每个文档中都含有特征词，也不能保证 $P(C_i/t)$ 所得的概率一定大。假设样本集有两个类别，C_1 类别有 200 个样例，共有 20 个样例含有特征词 t，C_2 类别有 20 个样例，每个样例中都含有特征词 t。通过计算，C_2 的 $P(C_i/t)$ 仅得 1/2，使得最终得到的 DFS(t) 不会很大，实际上对于 C_2 来说，特征词 t 会有很高分类效果。为了更具有说服力，本文选择两组数据集进行对比实验。第一组是

收稿日期：2017-03-21；**修回日期**：2017-05-16 **基金项目**：山东省科技发展计划资助项目(2014GGH201022)；山东省经信委软科学计划资助项目(2015EI010)；国家自然科学基金资助项目(71461027)

作者简介：付鑫(1992-)，男，山东寿光人，硕士研究生，主要研究方向为文本分类、信息抽取；王洪国，男(通信作者)，教授，博导，主要研究方向为电子政务、物流优化等(wang666666@126.com)；邵增珍，男，副教授，硕导，博士(后)，主要研究方向为智能计算、智能物流、大数据分析等；杜秋霞，女，硕士研究生，主要研究方向为信息抽取.

搜狗网站提供的公开均衡数据集,共 10 个类别,每个类别中含有 1 000 个样例,共 10 000 个样例;第二组是 Reuters-21578 不平衡数据集,包含 10 个类别,每个类别中含有的样例不均,个数存在明显差异。本文通过一定的实验验证,DFS 算法在均衡数据集下的准确度远大于不平衡数据集。DFS 另一个困境在于其未考虑特征词正负相关性的概念。特征词的正相关性也叫做特征词的积极性,是指在指定类别中存在的特征词对于此类别的影响;特征词的负相关也叫做特征词的消极性,是指在指定类别中不存在的特征词对于此类别的影响[15]。

如表 1 所示,六个文档包含三个类别 C_1、C_2、C_3, 其中,C_1 类别的所有文档都包含"猫"特征词,C_2、C_3 类别的所有文档都含有"狗"特征词,而"鱼"特征词占据 C_1 类别所有文档和 C_2 类别文档的一部分。分别计算三个特征词的 DFS 值得到:DFS(猫) = 1, DFS(鱼) = 0.7, DFS(狗) = 0.67。根据评分得到猫 > 鱼 > 狗排序,这是 DFS 评估函数得到的特征词重要性序列。但是根据特征词的分布情况可以看出,特征词"狗"对于类别 C_1 具有很好的分类效果(只要含有"狗"这个特征词的文档就不属于 C_1 类别,对于 C_1 的分类起关键性作用)。而 DFS 算法仅仅考虑特征词的整体作用,就会忽视掉很多局部有用的特征词。

表 1 特征词与类别

类别	特征词			类别	特征词
C_1	猫		鱼	C_2	狗
C_1	猫		鱼	C_3	狗
C_2		狗	鱼	C_3	狗

1.2 DFS 算法改进

针对上述 DFS 算法存在的两点不足,本文提出一定的优化策略,将 DFS 与卡方检测算法 CHI 结合,提出一种改进型特征选择算法(distinguishing feature selector-plus-CHI square),简称 DFS-sCHI 算法。CHI[4] 应用于评价两个变量之间的独立性程度,引入 CHI 作为评价特证词 t 与类别 C_i 的独立性程度。卡方值越大,说明特征词与此类别的相关性越大;反之,说明相关性越小。某个特征词 t 与某一类别 C_i 的卡方值如下所示:

$$\chi^2(t,C_i)=\frac{N(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \tag{2}$$

其中:N 表示训练样本的数量;A 表示类别 C_i 中出现特征词 t 的文档数量;B 表示非 C_i 类别中出现特征词 t 的文档数量;C 表示 C_i 类别中未出现特征词 t 的文档数量;D 表示非 C_i 类别中未出现特征词 t 的文档数量。

由于 DFS 考虑的是特征词的全局性,而 CHI 考虑的是特征词的局部性,所以 DFS-sCHI 可以看做是一种基于全局与局部特征的选择算法,既有特征词的全局属性又包含其局部属性。

1.2.1 算法描述

输入:训练样本集(不平衡数据集)。

输出:最优分类效果 top N 个特征词。

a)DFS 评估函数(式(1))计算每个特征词的 DFS 值,从大到小依次放入集合 $V(t)$ 中;

b)CHI 评估函数(式(2))计算 $V(t)$ 中各个特征词相对于各个类别的 CHI 值,得到特征词与类别的二维向量关系表;

c)选出二维向量每行中的最大值,公式 $F=AD-BC$ 计算此特征词与类别的相关性,整理结果如表 2 所示;

表 2 特征词评价指标

特征词序列	DFS 评分	CHI 评分	CHI 判断词性	最终评分
t_1				
t_2				
⋮				
t_n				

d)计算最终评分,由上到下遍历,若特征词 DFS 评分等于 1,则最终评分为 1 并记做特征词具有正相关性;反之,最终评分与特征词的词性根据 CHI 评分得到,最终评分取决于步骤 c)中最大值的绝对值,按由大到小的顺序排序,并将结果重新放入 $V(t)$ 中;

e)筛选特定个数特征词,根据特征词的词性与最终评分筛选特征词,每个类别所选特征词个数限定为 N,N 为所选特征词总个数/类别,筛选的结果放入集合 $S(t)$ 中。

f)算法结束,集合 $S(t)$ 即为降维后的列向量。

1.2.2 算法说明

a)计算预处理后所有特征词的 DFS 评分,若评分为 1,则说明此特征词对于某一类别有最高的分类精度,可直接将其提出并放入作为分类器的一维处理。

b)针对 DFS 算法的某些缺陷和不足,引入 CHI 评分机制,计算每个特征词特定于每个类别的详细评分,得到对应于每个类别的评分机制。

c)截取 CHI 评分机制的一部分 F 作为特征词特性的评价指标,若 $F<0$,说明特征词在其他类别中所占比例大,而在指定类别中所占比例小,即体现出特征词的负相关特性;若 $F>0$,说明特征词在指定类别中所占的比例大,即体现其正相关性。

d)选择特征词,若 DFS = 1,说明此特征词具有最高的分类精度、最高的优先级,将其提取到集合 $S(t)$ 中;其他的筛选过程根据步骤 b)c)特征词的 CHI 值和词性得到。

e)在不平衡数据集下增加提取特征词的准确度,需将提取的特征词均匀地分布于每个类别中,所以假定提取的特征词根据它的词性提取。每个类别中限定特征词为提取的总个数/总类别,即均匀分布于每个类别中,增加少样本类别的功能。

1.2.3 举例说明

以 Reuters-21578 不平衡数据集作为样本,共 10 个类别。整理步骤 a) ~ c)得到如表 3 所示表格(截取 15 个特征词)。

表 3 Reuters-21578 样本集评价指标

特征词序列	DFS 评分	CHI 评分			CHI 判断词性	最终评分
		C_1	C_2	C_3		
t_1	1.00	7.38	7.88	7.68	C_2:正相关	7.88
t_2	1.00	7.18	7.12	7.88	C_3:正相关	7.88
t_3	1.00	7.43	7.88	7.68	C_2:正相关	7.88
t_4	0.90	7.38	6.89	7.45	C_3:正相关	7.45
t_5	0.87	6.68	6.12	7.01	C_3:正相关	7.01
t_6	0.84	6.38	6.54	6.52	C_2:正相关	6.54
t_7	0.80	6.13	6.12	6.54	C_3:正相关	6.54
t_8	0.78	6.12	4.32	6.06	C_1:正相关	6.12
t_9	0.77	5.88	5.32	6.32	C_3:正相关	6.32
t_{10}	0.77	4.89	3.78	3.46	C_1:正相关	4.89
t_{11}	0.70	3.38	-4.12	3.68	C_2:负相关	4.12
t_{12}	0.65	3.45	4.45	3.23	C_2:正相关	4.45
t_{13}	0.60	-6.88	2.16	2.66	C_1:负相关	6.88
t_{14}	0.58	3.65	4.12	4.32	C_3:正相关	4.32
t_{15}	0.50	3.44	4.01	4.10	C_3:正相关	4.10

a)根据步骤 a)选出 DFS 值等于 1 的元素。

b)根据步骤 b) ~ d)计算最终评分,并对特征词重新排序,可以得到 t_{13} 在 t_5 之后、t_6 之前。根据步骤 e)保证每个类别中的特征词必须相同,筛选得到最终结果。算法结果如表 4 所示。

表 4 DFS 与 DFS-sCHI 算法结果

算法	选取的特征词	特征词类别
DFS	t_1、t_2、t_3、t_4、t_5、t_6、t_7、t_8、t_9	$C_1(1)$、$C_2(3)$、$C_3(5)$
DFS-sCHI	t_1、t_2、t_3、t_4、t_5、t_6、t_8、t_{10}、t_{13}	$C_1(3)$、$C_2(3)$、$C_3(3)$

2 实验及分析

2.1 实验数据集

实验选用三类数据集,第一类是搜狗网站提供的公开均衡数据集,包含 10 个类别,每个类别中含有 1 000 个样例,共 10 000个样例,类别依次是财经、房产、教育、科技、社会、时尚、时政、体育、游戏、娱乐;第二类是标准文档集——路透社文档集 Reuters-21578 不平衡数据集,包含 10 个类别,每个类别中含有的样例不均;第三类是实验室项目应用数据集——CNKI 白鹭摘要信息不平衡数据集,此数据集包含两个类别,第一个包含白鹭特征词的地理信息,第二个包含白鹭特征词的非地理信息。

2.2　分类器

本文选用人工神经网络作为分类器。人工神经网络(article neural network,ANN)[16]简称神经网络,是一种模仿生物神经网络结构和功能的数学与计算模型,用于对函数进行估计或近似。

现代神经网络是一种非线性统计数据建模工具,分为输入层、隐含层、输出层,输入层与隐含层之间及隐含层与输出层之间,通过神经进行连接,每个神经上有一个权值用于调优,训练的最终目的是计算得到所有神经上的最优权值。

训练过程如下:首先为每个神经上的权值赋初值;根据信号的正向传播及误差函数计算误差大小;根据误差判断是否需要对权值进行调优,若不需要,则为最优分类器,若需要,根据得到的 y 值计算输出单元的输出层权值增量与隐含层单元的权值增量;最后通过 n 次迭代得到最优神经网络分类器。

2.3　评价标准

为了评价分类效果,本文选用两类评价指标:macF1 和 micF1。其中,macF1 计算公式如下:

$$\text{macF1} = \sum_{k=1}^{C} F_k / C \tag{3}$$

$$F_k = \frac{2 \times P_k \times R_k}{P_k + R_k} \tag{4}$$

其中:准确率 P_k(precision)是指被正确分类的文档数除以被分类器识别为该类的文档数的商值;召回率 R_k(recall)是指被正确分类的文档数除以被测试文档总数的商值;F_k 是一种衡量分类总体效果的常用评估方法,其公式如式(4)所示。micF1 计算公式如下:

$$\text{micF1} = \frac{2 \times p \times r}{p + r}$$

其中:p 准确率、r 召回率表示范围为所有类别。

2.4　实验过程

实验采用 5 折交叉验证方法,将实验文本按照类别平均分成五份,其中四份为训练集,一份为测试集,每份轮流作为测试集,循环测试五次,取所有实验的平均值作为测试最终结果。

实验根据三组语料设置三组对比实验,macF1 和 micF1 作为分类精度好坏的评判指标。在搜狗均衡数据集上,分别选用 100、300、500、700、900、1 100、1 300 个特征作为向量空间维度,通过 DFS 与 DFS-sCHI 进行对比,得到的分类结果如图 1、2 所示。

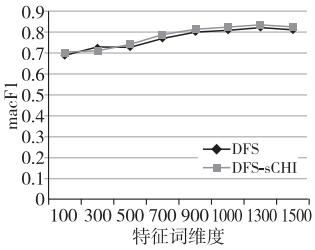


图 1　特征词维度与 macF1 值的关系

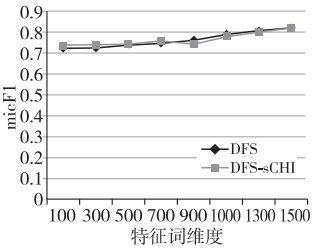


图 2　特征词维度与 micF1 的关系

由图 1、2 可以看出,均衡数据集下,DFS 与 DFS-sCHI 算法的 macF1 与 micF1 值相差不大。原因在于均衡样本集下,各个类别样本数相同,DFS 算法在选取特征词时,极小概率会出现在某些类别中选取不到的情况,极大概率下特征词均匀分布;而 DFS-sCHI 算法的最终目标是根据特征词重要性均分特征词,两者在均衡样本下的结果殊途同归。所以两者的 macF1 与 micF1 值相差不大且都有很高的分类精度。由此得到均衡数据集下 DFS 与 DFS-sCHI 算法的表现效果基本相同。

在路透社文档集 Reuters-21578 不平衡数据集下,同样选用 100、300、500、700、900、1 100、1 300 个特征作为向量空间维度,通过 DFS 与 DFS-sCHI 进行对比,得到的结果如图 3、4 所示。

由图 3、4 可以看出,不平衡数据集下,DFS 与 DFS-sCHI 算法得到的分类精度有所不同,无论是 macF1 值还是 micF1 值,DFS-sCHI 算法都有显著提高。原因在于不平衡数据集下,应用 DFS 算法有极大概率会忽视小类别样本,即选取的特征词未存在于小类别样本中,从而导致很多高效分类的特征词被算法所忽略,影响分类精度。DFS-sCHI 算法通过引入 CHI,二次回收被 DFS 算法忽视而分类效果较高的词,所以分类精度较 DFS 有一定提高。由此得到 DFS-sCHI 算法在不均衡数据集下比传统的 DFS 算法拥有更好的分类效果。

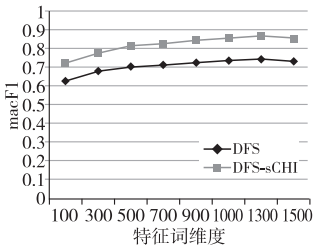


图 3　特征词维度与 macF1 值的关系

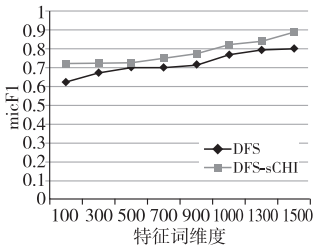


图 4　特征词维度与 micF1 的关系

在 CNKI 白鹭摘要信息不平衡数据集下,CNKI 白鹭摘要信息样本集是实验室为获取鸟情数据自主采集的样本集,然而 CNKI 采集的数据未必就与鸟情数据有关(含有白鹭的词未必就是鸟情数据),需要对其进行分类处理,选择其中的鸟情样本。本文所爬取的样本集是二元不平衡样本且类别相差较大,因此将两种算法应用于此样本集上进行对比实验,得到的结果如图 5、6 所示。

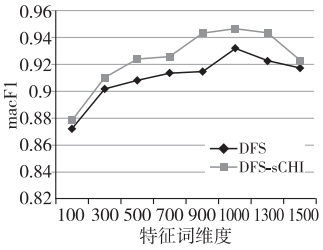


图 5　特征词维度与 macF1 值的关系

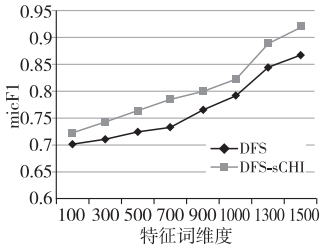


图 6　特征词维度与 micF1 的关系

由图 5、6 可以看出,对于不平衡数据集下的二元分类问题,DFS-sCHI 算法相较于 DFS macF1 与 micF1 值有显著提高,且选取的特征词维数越多,精度越高。原因在于二元分类问题同样类似于多元分类,也会出现特征词分布不均的情况。由此得到 DFS-sCHI 算法在二元不平衡数据集下比传统的 DFS 算法拥有更好的分类效果。

通过三组对比实验可以看出,不管是在二元分类还是多元分类问题,在不平衡数据集下,DFS-sCHI 的分类效果远大于 DFS;而在均衡数据集下,两种算法的精度相差不大。

3　结束语

本文从 DFS 算法在不平衡数据集上分类精度下降且未考虑负相关特征词对分类精度的影响这两方面因素考虑,结合局部特征选择算法 CHI 的优势,提出 DFS-sCHI 算法,强调特征词类别属性和两面性,可以有效地提高 DFS 算法在不平衡样本下的分类精度。算法将 DFS 与 CHI 结合,必然会加大算法的时空复杂度,如何在复杂度与精度之间寻求一个向量空间模型维度最优解,是下一步要解决的问题[17]。

参考文献:

[1] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[2] Yang Yiming, Pedersen J O. A comparative study on feature selection in text categorization[C]//Proc of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 1997: 412-420.

[3] Lee C, Lee G G. Information gain and divergence-based feature selection for machine learning-based text categorization[J]. Information Processing & Management, 2006, 42(1): 155-165.

(下转第 1969 页)

的差分进化粒子滤波的实时性能最佳。改进算法在 8 核下的加速比如表 3 所示。

表 3　改进算法在 8 核下的加速比

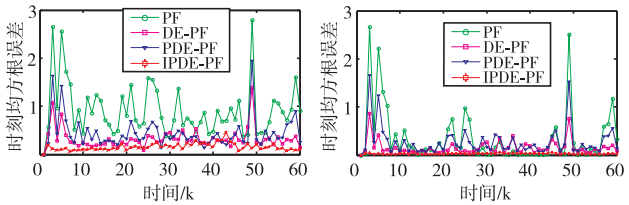
粒子数	运行时间/s		加速比
	CPU	GPU	
1 000	0.022	0.01	2.2
2 000	0.084	0.03	2.8
4 000	0.2	0.05	4
5 000	0.308	0.07	4.4
6 000	0.36	0.08	4.5
8 000	0.522	0.09	5.8
10 000	1.6	0.16	10

实验 2　均方根误差的比较

对算法进行 $R_{MC}=200$ 次独立蒙特卡洛仿真,并定义时刻 k 的均方根误差为

$$L_k^{\mathrm{RMSE}}=\sqrt{\frac{1}{R_{MC}}\sum_{j=1}^{R_{MC}}(x_{k,j}-\bar{x}_{k,j})^2} \tag{15}$$

其中:$x_{k,j}$和 $\bar{x}_{k,j}$分别表示第 j 次仿真中 k 时刻的实际状态和预测状态。量测噪声 $v_k\sim N(0,0.000\ 01)$。图 10 给出粒子数 $N=100$ 和 $N=200$ 两种设置下四种算法的时刻均方根误差对比。从图 10 可看出,IPDE-PF 的时刻均方根误差最小,说明其估计精度最高,且随着粒子数的减少估计精度仍优于另外三种算法。从图可知,采用并行前缀求和优化算法的差分进化粒子滤波误差明显小于标准粒子滤波算法,每种方法优化后的粒子滤波精度存在较大差别,IPDE-PF 的滤波精度最高;PDE-PF 的滤波精度高于 DE-PF,但算法运行时间最长,由此可见共享内存访问 bank 冲突对最终的优化效果具有显著影响;本文 IPDE-PF 算法在粒子数较少量测噪声较大的情况下,跟踪精度均好于其他三种算法。



(a)粒子数为100时均方根误差对比　(b)粒子数为200时均方根误差对比

图 10　不同粒子数的均方根误差对比

5　结束语

通过分析已有智能优化粒子滤波存在的问题,本文使用零 bank 冲突规约算法优化差分进化粒子滤波,通过采用带填充寻址方式,解决共享内存中内存访问时出现的 bank 冲突问题,提高了规约算法的执行效率,利用改进的并行前缀求和算法来优化差分进化粒子滤波,大大提高了差分进化粒子滤波算法在 GPU 上的执行效率。实验结果表明,所提出的改进规约的差分进化粒子滤波算法有效提高了实时性,能适应一般系统的应用。

下一步工作中将进一步对并行差分进化粒子滤波算法进行优化,提高可并行部分所占的比例,从而进一步提高算法整体的实时性能。

参考文献:

[1] Sileshi B G, Oliver J, Toledo R, *et al*. Particle filter SLAM on FPGA: a case study on Robot@ Factory competition[C]//Proc of the 2nd Iberian Robotics Conference. Berlin: Springer International Publishing, 2016.

[2] Pardal P C P M, Kuga H K, De Moraes R V. The particle filter sample impoverishment problem in the orbit determination application [M]//Mathematical Problems in Engineering. 2015: 1-9.

[3] Malarvezhi P, Kumar R. Particle filter with novel resampling algorithm: a diversity enhanced particle filter[J]. Wireless Personal Communications, 2015, 84(4): 3171-3177.

[4] 余春超, 杨智雄, 夏宗泽, 等. 采用 GPU 并行架构的基于互信息和粒子群算法的异源图像配准[J]. 红外技术, 2016, 38(11): 938-946.

[5] Li Hongwei, Wang Jun, Su Hongtao. Improved particle filter based on differential evolution[J]. Electronics Letters, 2011, 47(19): 1078-1079.

[6] 孙伟平, 向杰, 陈加忠, 等. 基于 GPU 的粒子滤波并行算法[J]. 华中科技大学学报: 自然科学版, 2011,39(5): 63-66.

[7] 刘光敏, 陈庆奎, 赵海燕, 等. 一种 GPU 加速的粒子滤波算法[J]. 计算机科学, 2014(Z11).

[8] Murray L M, Lee A, Jacob P E. Parallel resampling in the particle filter[J]. Journal of Computational & Graphical Statistics, 2013,25(3):789-805.

[9] 赵嵩, 徐彦, 曹海旺, 等. GPU 并行实现多特征融合粒子滤波目标跟踪算法[J]. 微电子学与计算机, 2015,32(9): 153-156,160.

[10] 刘伟, 孟朝晖, 薛东伟. 基于 CUDA 与粒子滤波的多特征融合视频目标跟踪算法[J]. 计算机系统应用, 2013, 22(11): 123-128.

[11] 武勇, 王俊, 曹运合, 等. 基于二次预测的粒子滤波算法[J]. 吉林大学学报: 工学版, 2015, 45(5): 1696-1701.

[12] 张硕, 何发智, 周毅, 等. 基于自适应线程束的 GPU 并行粒子群优化算法[J]. 计算机应用, 2016, 36(12): 3274-3279.

[13] 余莹, 李肯立, 郑光勇. 一种基于 GPU 集群的深度优先并行算法设计与实现[J]. 计算机科学, 2015, 42(1): 82-85.

[14] Padua D. All prefix sums [M]// Encyclopedia of Parallel Computing. 2011.

[15] Nakano K. An optimal parallel prefix-sums algorithm on the memory machine models for GPUs[C]//Proc of the 12th International Conference on Algorithms and Architectures for Parallel Processing. 2012: 99-113

(上接第 1949 页)

[4] Chen Y T, Chen Mengchang. Using Chi-square statistics to measure similarities for text categorization[J]. Expert Systems with Applications, 2011, 38(4): 3085-3090.

[5] Liu Huawen, Sun Jigui, Liu Lei, *et al*. Feature selection with dynamic mutual information[J]. Pattern Recognition, 2009, 42(7): 1330-1339.

[6] Shang Wenqian, Huang Houkuan, Zhu Haibin, *et al*. A novel feature selection algorithm for text categorization[J]. Expert Systems with Applications, 2007, 33(1): 1-5.

[7] Ogura H, Amano H, Kondo M. Feature selection with a measure of deviations from Poisson in text categorization[J]. Expert Systems with Applications, 2009, 36(3): 6826-6832.

[8] Yang Jieming, Qu Zhaoyang, Liu Zhiying. Improved feature-selection method considering the imbalance problem in text categorization[J]. The Scientific World Journal, 2014, 2014(5):articleID 625342.

[9] 闫健卓, 李鹏英, 方丽英, 等. 基于 χ^2 统计的改进文本特征选择方法[J]. 计算机工程与设计, 2016, 37(5): 1391-1394.

[10] 任永功, 杨雪, 杨荣杰, 等. 基于信息增益特征关联树的文本特征选择算法[J]. 计算机科学, 2013, 40(10): 252-256.

[11] 尤鸣宇, 陈燕, 李国正. 不均衡问题中的特征选择新算法: Im-IG[J]. 山东大学学报: 工学版, 2010, 40(5): 123-128.

[12] 徐燕, 李锦涛, 王斌, 等. 不均衡数据集上文本分类的特征选择研究[J]. 计算机研究与发展, 2007,44(S1):58-62.

[13] 张玉芳, 王勇, 熊忠阳, 等. 不平衡数据集上的文本分类特征选择新方法[J]. 计算机应用研究, 2011, 28(12):4532-4534.

[14] Uysal A K, Gunal S. A novel probabilistic feature selection method for text classification[J]. Knowledge-Based Systems, 2012, 36(12): 226-235.

[15] Ogura H, Amano H, Kondo M. Comparison of metrics for feature selection in imbalanced text classification[J]. Expert Systems with Applications, 2011, 38(5): 4978-4989.

[16] Fausett LV. Fundamentals of neural networks: architectures, algorithms, and applications[M]. New Jersey: Prentice-Hall Inc, 1994.

[17] Pietramala A, Policicchio V L, Rullo P. Automatic filtering of valuable features for text categorization[C]// Proc of International Conference on Advanced Data Mining and Applications. Berlin: Springer, 2012: 284-295.