基于信息增益的多标签特征选择算法

李 玲[1] 刘华文[1,2] 徐晓丹[1] 赵建民[1]

(浙江师范大学数理与信息工程学院 金华 321004)[1] (中国科学院数学与系统科学研究院 北京 100055)[2]

摘 要 多标签特征选择是一种提高多标签分类器性能的技术。针对目前这类技术在给出合理特征子集合时无法同时兼顾计算复杂度和标签间的相关性的问题,提出一种基于信息增益的多标签分类算法。该算法假设特征之间相互独立,首先使用单个特征与整个标签集合之间的信息增益来度量这两者的关联程度,再根据阈值删除不相关的特征以得到最优特征子集合。实验表明,该算法能有效地提高多标签分类器的分类性能。

关键词 数据挖掘,多标签分类,特征选择,信息增益

中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.012

Multi-label Feature Selection Algorithm Based on Information Gain

LI Ling[1] LIU Hua-wen[1,2] XU Xiao-dan[1] ZHAO Jian-min[1]

(College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004, China)[1]

(Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100055, China)[2]

Abstract Multi-label feature selection is a kind of technology which is used to improve the performance of multi-label classifiers. However, the existing multi-label feature selection methods fail to make a tradeoff between the possible dependence among the labels and computational complexity in the process of obtaining reasonable feature subsets. Therefore, a novel multi-label feature selection algorithm based on information gain was proposed in the essay. It assumes that the features are independent with each other. The proposed method firstly uses information gain between a single feature and a set of labels to measure their correlation degree, and then removes the irrelevant and redundant features according to a threshold value. The experimental results show that the proposed algorithm can more effectively promote the performance of multi-label classifiers.

Keywords Data mining, Multi-label learning, Feature selection, Information gain

1 引言

多标签分类是数据挖掘领域的研究热点之一,并已在蛋白质功能分类[1]、文本分类[2]、语义场景分类[3]等领域取得广泛应用。多标签分类是针对多标签数据的特点,获取相应的分类模型,并依此判断未知数据的类别的过程[4]。与传统的单标签分类一样,多标签分类同样面临维灾难[5]问题。为了解决这个问题,通常的做法是对多标签数据实施特征选择操作,以降低高维性所引起的不利影响。

特征选择是依据给定的评价标准,选择一个最能保持数据原始特性的最优特征子集的过程[6]。迄今为止,已有多种特征选择算法被提出。其中,多标签特征选择是一种适用于多标签数据的特征选择技术。目前,这类技术较少[7]。传统的处理方式是首先使用转化方法将多标签数据转化为单标签数据,再直接应用成熟的单标签特征选择技术,以解决多标签特征选择问题。例如,Lee等人[8]提出了一种基于近似信息增益的算法。该算法采用近似信息增益度量分别度量特征与标签集合和特征与已选特征集合的关联程度,每次选择与标签集合关联最大而与已选特征集合关联最小的特征。该算法只能获得一个特征排名序列。

为此,Spolao等人[9]提出了一种能直接给出特征子集合的算法。该算法首先针对每个特征,分别使用评价标准度量增益度量特征与标签集合中的每个标签的关联程度;再对它们实施平均化操作;然后以均值评价该特征的重要程度;最后,根据阈值剔除不重要的特征。若评价标准是ReliefF,则阈值为0.01;若评价标准是信息增益,则阈值为0.1。张振海等人[10]提出了一种基于信息增益的多标签特征选择算法,该算法以特征与所有单个标签间的信息增益作为特征与标签集合的相关性量度,再根据阈值不相关的特征来获得一个最优的特征子集。虽然这两种方法都能获得最优特征子集合,但它们并没有考虑标签间可能存在的关联。实际上,这种关联是一种可以帮助分类器准确地预测和分类的重要信息。

到稿日期:2014-06-20 返修日期:2014-10-25 本文受国家自然科学基金(61100119,61272130,61272468,61170108,61170109),模式识别国家重点实验室开放课题基金(201204214),中国博士后基金(2013M530072),浙江省自然科学基金项目(LY14F020012),浙江省教育厅项目(Y201328291)资助。

李 玲(1989-),女,硕士,主要研究方向为数据挖掘、模式识别、多标签分类,E-mail:human1205@163.com;刘华文(1977-),男,博士,副教授,主要研究方向为数据挖掘、模式识别、多标签分类;徐晓丹(1977-),女,硕士,讲师,主要研究方向为数据挖掘、自然语言处理;赵建民(1950-),男,硕士,教授,主要研究方向为数据挖掘、模式识别、云计算、多标签分类。

针对上述算法的不足之处，本文提出一种基于信息增益的多标签特征选择算法(Multi-label Feature Selection Algorithm Based on Information Gain,FSIG)。该算法能在特征选择过程中充分利用标签间的相关性并获得最优特征子集合。它首先用这个特征与标签集合的信息增益来度量特征的重要程度，再根据设定的阈值删除不重要的特征，以达到特征选择的目的。实验结果表明，该算法能有效地提高多标签器的性能。

2 相关工作

与传统的特征选择一样，多标签特征选择也可大致分为3类：封装式、嵌入式和过滤式。封装式方法将特征选择算法作为分类算法的一部分，并直接使用分类性能作为评价特征的标准[6]。例如，张永波等人[11]提出了一种以平均分类精度作为特征子集的适应函数，以模拟退火策略来寻找最优特征子集的多标签特征算法。之后，邵欢等人[12]在该算法的基础上，提出了一种结合模拟退火算法、遗传算法和贪心算法这3种算法的混合多标签特征选择算法，以进一步优化最优特征子集。

嵌入式方法，如尤鸣宇等人[13]提出的算法，也是利用分类器来进行特征选择操作。与封装式方法不同，它同时进行特征选择过程与学习训练过程，并在学习训练过程中找到最优特征子集合[6]。这两类方法都依赖于分类器，且存在效率低、占用内存大等不足。

与前两类方法不同，过滤式方法独立于分类器，并具有操作简单、速度快等优点[6]。过滤式方法就是根据一定的标准，在原始特征空间中寻找一个最能保持数据原始特性的特征或特征集合[14]。近年来，已有不少这类算法被提出。例如，Doquire等人[15]提出了一种基于信息增益的多标签特征选择算法，该算法先使用PPT策略将多标签数据转化为单标签数据，再使用信息增益来度量特征与标签的关联程度，然后结合前向搜索策略以选择最优特征。该算法能在一定程度上考虑到标签间的相关性，但是无法给出一个合理的特征子集合。

3 基于信息增益的多标签特征选择算法

3.1 信息增益

在信息论中，信息增益(Information Gain)是一种量化随机变量 S 和 Z 的关联程度的量度，其值可由式(1)[16]计算：

$$IG(S;Z)=H(S)+H(Z)-H(S,Z) \tag{1}$$

其中，$H(S)$表示变量 S 的信息熵，$H(Z)$表示变量 Z 的信息熵，$H(S,Z)$表示变量 S 和 Z 的联合熵。

3.2 交互信息

在信息论中，交互信息(Interaction Information)是一种概括性的信息量。给定一个集合 T，则 $I(T)$表示集合 T 的交互信息，其计算公式[17]如下：

$$I(T)=-\sum_{X\in T'}(-1)^{|X|}H(X) \tag{2}$$

其中，T'是由集合 T 的所有子集合组成的集合，$|X|$表示 X 中的变量个数，$H(X)$表示变量 X 的信息熵。

若 $T=\{t_1,t_2\}$，t_1 和 t_2 是两个随机变量，则

$$T'=\{\phi,t_1,t_2,\{t_1,t_2\}\} \tag{3}$$

$$I(T)=H(t_1)+H(t_2)-H(t_1,t_2) \tag{4}$$

即 $$I(\{t_1,t_2\})=IG(t_1;t_2) \tag{5}$$

若 $T=\{t,T_1\}$，t 是一个随机变量，T_1 是一个变量集合，$T_1'=\{t_1,t_2,t_3,\cdots,t_n\}$是由集合 T_1 的所有子集合组成的集合，n 是集合 T_1 的子集总数，则

$$T'=T_1'\cup\{\{t\},\{t,t_1\},\{t,t_2\},\cdots,\{t,t_n\}\} \tag{6}$$

$$\begin{aligned}I(T)&=-\sum_{X\in T'}(-1)^{|X|}H(X)\\&=-\sum_{i=1}^{n}(-1)^{|t_i|}H(t_i)-\sum_{i=1}^{n}(-1)^{|\{t,t_i\}|}H(\{t,t_i\})\\&=\sum_{i=1}^{n}(-1)^{|t_i|}[H(\{t,t_i\})-H(t_i)]\end{aligned} \tag{7}$$

3.3 特征与标签集合的相关性

给定一个特征 f 和标签集合 Y，则信息增益 $IG(f;Y)$可表征特征 f 和标签集合 Y 的关联程度。它可分解为多个变量的交互信息[18]，如式(8)所示：

$$\begin{aligned}IG(f;Y)&=\sum_{i=1}^{m}\sum_{y\in Y_i}I(\{f,y\})\\&=\sum_{y\in Y_1}I(\{f,y\})+\sum_{i=2}^{m}\sum_{y\in Y_i}I(\{f,y\})\\&=\sum_{i=1}^{m}I(\{f,y_i\})+\sum_{i=2}^{m}\sum_{y\in Y_i}I(\{f,y\})\end{aligned} \tag{8}$$

其中，m 是标签个数，Y_i 是由集合 Y 的含有 i 个元素的子集组成的集合。根据式(5)和式(7)可将式(8)改写为：

$$IG(f;Y)=\sum_{i=1}^{m}IG(f;y_i)+\sum_{i=2}^{m}\sum_{y\in Y_i}\sum_{i=1}^{n}(-1)^{|y_1|}\cdot[H(\{f,y_1\})-H(y_1)] \tag{9}$$

其中，n 是集合 y 的子集总数，y_i 是由集合 y 的第 i 个子集。因此，在不破坏原始标签关联结构的情况下，$IG(f;Y)$不仅能表征特征 f 与标签集合 Y 中的所有标签的关联程度，还能表征特征 y 对标签集合 Y 中的所有标签组合的影响。

3.4 FSIG算法描述

本文使用特征与标签集合的信息增益 $IG(f;Y)$来评价特征的重要程度。事实上，各个特征与标签集合的信息增益不一定在同一衡量范围内。为了便于比较，本文对信息增益 $IG(f;Y)$进行了归一化处理，具体操作如式(10)所示：

$$SU(f,Y)=\frac{2*IG(f;Y)}{H(f)+H(Y)} \tag{10}$$

其中，$SU(f,Y)\in[0,1]$，$SU(f,Y)$越大，表示 f 和 Y 的关联性越大。当 $SU(f,Y)=0$ 时，表示 f 和 Y 相互独立；当 $SU(f,Y)=1$时，表示可由其中任一个量确定另一个量，即已知 f 可以确定 Y，同样，已知 Y 可以确定 f。

总体上，FSIG算法首先使用信息增益分别度量每个特征与标签集合的关联程度；再对它们的信息增益实施归一化操作，以使其均在同一度量范围；然后计算所有特征的归一化信息增益的均值，并以该均值作为阈值；最后根据阈值删除不相关的特征得到最优特征子集合。该算法的具体实现如表1所列。

表1 基于信息增益的多标签特征选择算法

```
算法 BF=FSIG(F,Y)
输入：特征集合 F={f1,f2,f3,…,fn}
      标签集合 Y={y1,y2,y3,…,yq}
输出：最优特征子集 BF
(1)IGS=φ;
(2)for i=1 to n do
(3)  IG(fi;Y)=H(fi)+H(Y)-H(fi,Y);
(4)  SU(fi,Y)=2*IG(fi;Y)/(H(fi)+H(Y));
(5)  IGSi=SU(fi,Y);
(6)  IGS=IGS∪IGSi;
(7)endfor
(8)μ=(1/n)*Σ_{i=1}^{n} IGSi
(9)BF=F;
(10)for i=1 to n do
(11)  if IGSi<μ then BF=BF-{fi};
(12)endfor
```

4 实验

4.1 实验设置

为了验证本文算法的有效性，本实验将其与 MLFSIE 算法[10]、RF-BR 算法[9]、AMI 算法[8]、Avg Relief 算法[19]进行对比，采用 ML-RBF[20]和 InsDif[21]这两种多标签分类器对特征选择后的数据集进行验证。MLFSIE 算法和 RF-BR 算法都是能获得最优特征子集的过滤式方法。AMI 算法和 Avg Relief 算法都无法获得最优特征子集。为了保证算法之间具有可比性，它们保留的特征数目与 FSIG 算法特征选择后的特征数目相同。实验中，ML-RBF 分类器中的缩放因子 μ 的值为 1，InsDif 分类器中的聚类数比率 M 取值为 0.02。实验过程首先使用特征选择算法对数据集进行降维，然后再使用分类器对降维后的数据集以十字交叉方法进行验证。

4.2 实验数据集

采用 cal500、emotion、enron 和 scene 4 种不同的公共数据集，相关信息如表 2 所列。

表 2 实验数据集的相关信息

数据集	cal500	emotion	enron	scene
样本数	502	593	1702	2407
特征数	68	72	1001	294
标签数	174	6	53	6
标签基数	26.044	1.869	3.378	1.074
标签密度	0.150	0.311	0.064	0.179
标签组合数	502	27	753	15

4.3 实验结果及分析

本实验采用 5 种评价指标[22]，即平均精度(Average precision)、结果覆盖长度(Coverage)、汉明损失(Hamming Loss)、1-错误率(One-error)和排名损失(Ranking Loss)分别从不同的角度衡量分类性能的好坏，并直接体现在数值的大小上。这些指标在实验数据集上的取值分别如表 3 至表 12 所列，其中，粗体表示最优值，"↑"表示取值越大分类性能越好，而"↓"表示取值越小分类性能越好。

(1)平均精度

平均精度用于刻画分类器的标签预测准确程度[22]。该指标取值越大，表示分类器的性能越好。从表 3 和表 4 可知，在这 4 种不同数据集上，FSIG 算法总是略优于其它 4 种对比算法。因此，从平均精度方面看，FSIG 算法略优于其它 4 种对比算法。

表 3 特征选择后 MLRBF 分类器的平均精度比较(↑)

数据集	cal500	emotion	enron	scene
FSIG	**0.355577**	**0.774727**	**0.707707**	**0.858521**
MLFSIE	0.335642	0.630099	0.691114	0.826403
RF-BR	0.341497	0.751342	0.692735	0.830722
AMI	0.350808	0.749915	0.702818	0.852264
Avg Relief	0.347900	0.749228	0.599251	0.818809

表 4 特征选择后 InsDif 分类器的平均精度比较(↑)

数据集	cal500	emotion	enron	scene
FSIG	**0.488737**	**0.796219**	**0.700694**	**0.871389**
MLFSIE	0.453262	0.512702	0.678253	0.848063
RF-BR	0.462398	0.759991	0.686861	0.856441
AMI	0.485340	0.757128	0.690680	0.870017
Avg Relief	0.476881	0.755829	0.576069	0.850188

(2)结果覆盖长度

结果覆盖长度用于考察样本的标签排序序列覆盖隶属于样本的所有标签所需要的搜索深度情况[22]。该指标取值越小，表示分类器的性能越好。从表 5 可知，在 cal500 数据集上，FSIG 算法只略优于 AMI 算法；在 emotion 和 enron 数据集上，FSIG 算法略优于其它 4 种算法；在 scene 数据集上，FSIG 算法略优于 MLFSIE 算法、RF-BR 算法和 Avg Relief 算法。从表 6 可知，在 cal500 数据集上，FSIG 算法略优于 MLFSIE 算法、RF-BR 算法和 Avg Relief 算法；在 emotion 数据集上，FSIG 算法略优于其它 4 种算法；在 enron 数据集上，FSIG 算法略优于 MLFSIE 算法、AMI 算法和 Avg Relief 算法；在 scene 数据集上，FSIGL 算法略优于 MLFSIE 算法、RF-BR 算法和 Avg Relief 算法。然而在不同的数据集和不同的分类器上，这 5 种算法的优劣排序不断在变化，但 FSIG 算法的平均排名是最优的。因此，从结果覆盖长度方面看，FSIG 算法略优于其它 4 种对比算法。

表 5 特征选择后 MLRBF 分类器的结果覆盖长度比较(↓)

数据集	cal500	emotion	enron	scene
FSIG	164.414000	**1.866102**	**13.771895**	0.526667
MLFSIE	163.698000	2.722034	14.313072	0.616250
RF-BR	164.264000	2.020339	14.527451	0.582917
AMI	164.472000	2.023729	14.114379	**0.519583**
Avg Relief	**163.510000**	2.023729	16.562745	0.616667

表 6 特征选择后 InsDif 分类器的结果覆盖长度比较(↓)

数据集	cal500	emotion	enron	scene
FSIG	145.408000	**1.788136**	14.264706	0.480833
MLFSIE	146.664000	3.384746	14.625294	0.547083
RF-BR	145.796000	2.006780	**14.085882**	0.506667
AMI	**144.064000**	2.027119	14.405882	**0.455417**
Avg Relief	146.602000	2.001695	17.977647	0.534167

(3)汉明损失

汉明损失适用于属于该样本的标签未出现在该类标签集中而不属于的却出现的情况[22]。该指标取值越小，表示分类器的性能越好。从表 7 和表 8 可知，在 cal500、emotion 和 enron 数据上，FSIG 算法略优于其他 4 种算法；在 scene 数据集上，FSIG 算法仅略优于 MLFSIE 算法、RF-BR 算法。因此，从汉明损失方面看，FSIG 算法略优于其他 4 种算法。

表 7 特征选择后 MLRBF 分类器的汉明损失比较(↓)

数据集	cal500	emotion	enron	scene
FSIG	**0.885287**	**0.217514**	**0.044790**	0.097431
MLFSIE	0.890080	0.308475	0.045308	0.109861
RF-BR	0.892448	0.238418	0.046233	0.111667
AMI	0.892644	0.236441	0.045234	**0.095000**
Avg Relief	0.892264	0.240113	0.054458	0.111667

表 8 特征选择后 InsDif 分类器的汉明损失比较(↓)

数据集	cal500	emotion	enron	scene
FSIG	**0.857299**	**0.204802**	**0.045305**	0.085208
MLFSIE	0.857966	0.318644	0.046093	0.096667
RF-BR	0.867977	0.235876	0.046271	0.093889
AMI	0.867724	0.230791	0.045727	**0.084167**
Avg Relief	0.868207	0.233051	0.054473	0.097778

(4)1-错误率

1-错误率简单统计了在样本预测标签集中，排序靠前但又不应该属于这个集合的标签个数[22]。由表 9 和表 10 可

知，在 emotion、enron 和 scene 数据集上，FSIG 算法略优于其它 4 种算法。在 cal500 数据上，FSIG 算法的表现略差些。当使用 MLRBF 分类器时，FSIG 算法仅略优于 MLFSIE 算法、RF-BR 算法和 Avg Relief 算法；当使用 InsDif 分类器时，FSIG 算法仅略优于 MLFSIE 算法、AMI 算法和 Avg Relief 算法。因此，从 1-错误率方面看，FSIG 算法略优于其它 4 种算法。

表 9　特征选择后 MLRBF 分类器的 1-错误率比较(↓)

数据集	cal500	emotion	enron	scene
FSIG	0.434000	**0.327119**	**0.228758**	**0.228750**
MLFSIE	0.444000	0.510169	0.247712	0.286667
RF-BR	0.452000	0.342373	0.242484	0.283333
AMI	**0.408000**	0.349153	0.243137	0.250417
Avg Relief	0.462000	0.350847	0.372549	0.305833

表 10　特征选择后 InsDif 分类器的 1-错误率比较(↓)

数据集	cal500	emotion	enron	scene
FSIG	0.178000	**0.291525**	**0.217647**	**0.209583**
MLFSIE	0.184000	0.672881	0.252353	0.249583
RF-BR	**0.168000**	0.325424	0.236471	0.239167
AMI	0.186000	0.323729	0.240588	0.218750
Avg Relief	0.180000	0.350847	0.382941	0.248750

(5)排名损失

排名损失用于描述所有标签排序的出错程度[22]，由表 11 可知，在 cal500、emotion 和 enron 数据集上，FSIG 算法略优于其它 4 种算法；在 scene 数据集上，FSIG 算法仅略差于 AMI 算法。从表 12 可知，在 cal500 和 emotion 数据集上，FSIG 算法略优于其它 4 种算法；在 enron 数据集上，FSIG 算法仅略差于 RF-BR 算法；在 scene 数据集上，FSIG 算法仅略差于 AMI 算法。因此，从排名损失来看，FSIG 算法略优于其它 4 种算法。

表 11　特征选择后 MLRBF 分类器的排名损失比较(↓)

数据集	cal500	emotion	enron	scene
FSIG	**0.333400**	**0.183856**	**0.088259**	0.087145
MLFSIE	0.358037	0.369021	0.093707	0.105132
RF-BR	0.343126	0.212622	0.095487	0.098167
AMI	0.339775	0.213508	0.090863	**0.085948**
Avg Relief	0.337115	0.214143	0.110210	0.104938

表 12　特征选择后 InsDif 分类器的排名损失比较(↓)

数据集	cal500	emotion	enron	scene
FSIG	**0.205101**	**0.167637**	0.089047	0.077219
MLFSIE	0.224012	0.371186	0.091009	0.091267
RF-BR	0.217131	0.210885	**0.088102**	0.082698
AMI	0.205147	0.216017	0.089386	**0.072770**
Avg Relief	0.211429	0.213197	0.110030	0.088052

虽然，在不同分类器和不同数据集上，这 5 种方法的优劣排序在不断变化，但总体上 FSIG 算法略优于其它 4 种方法。在平均精度方面，FSIG 算法的表现尤为突出。

结束语　本文以特征与标签集合的信息增益作为评价关联程度的准则，提出了一种基于信息增益的多标签特征选择算法，简称 FSIG。FSIG 算法的主要特点是利用标签之间已存在的相关性进行特征选择。因此，FSIG 算法在处理多标签数据时，可以完整地保留原始标签间的相关性。实际上，这种相关性可以提高多标签分类器性能。实验结果表明，与 4 种常见的多标签特征选择算法相比，FSIG 算法可以获得较好的分类结果。

参 考 文 献

[1] Elisseeff A, Weston J. A kernel method for multi-labelled classification[C]// NIPS. 2001: 681-687

[2] Lewis D D, Yang Y, et al. A new benchmark collection for text categorization research[J]. Journal of Machine Learning Research, 2004, 5: 361-397

[3] Boutell M R, Luo J, et al. Learning multi-label scene clssifycation[J]. Pattern Recognition, 2004, 37(9): 1757-1771

[4] Tsoumakas G, Katakis I, et al. Mining multi-label data[M]// Data Mining and Knowledge Discovery Handbook. New York: Springer US, 2010: 667-685

[5] Liu Hua-wen, Li Min-shuo, et al. An effective feature selection method using dynamic information criterion[J]. In Artificial Intelligence and Computational Intelligence, 2011, 7002: 450-455

[6] 刘华文. 基于信息熵的特征选择算法研究[D]. 吉林：吉林大学，2010
Liu Hua-wen. A Study on Feature Selection Algorithms using Information Entropy[D]. Jilin: Jilin University, 2010

[7] Doquire G, Verleysen M. Mutual information-based feature selection for multilabel classification[J]. Neurocomputing, 2013, 122: 148-155

[8] Lee J, Lim H, Kim D W. Approximating mutual informatioon for multi-label feature selection[J]. Electronics Letters, 2012, 48(15): 929-930

[9] Spolaör N, Cherman E A, et al. Filter approach feature selection methods to support multi-label learning based on reliefF and information gain[M]// Advances in Artificial Intelligence-SBIA 2012. Springer Berlin Heidelberg, 2012: 72-81

[10] 张振海，李士宁，等. 一类基于信息熵的多标签特征算法[J]. 计算机研究与发展，2013，50(6)：1177-1184
Zhang Zhen-hai, Li Shi-ning, et al. Multi-Label Feature Selection Algorithm Based on Information Entropy[J]. Journal of Computer Research and Development, 2013, 50(6): 1177-1184

[11] 张永波，游录金，等. 基于模拟退火的多标记数据特征选择[J]. 计算机工程与设计，2011，32(7)：2494-2500
Zhang Yong-bo, You Lu-jin, et al. Feature selection for multi-label data by using simulated annealing[J]. Computer Engineering and Design, 2011, 32(7): 2494-2500

[12] Shao huan, Li Guo-zheng, et al. Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine[J]. Science China Information Sciences, 2011, 56(5): 1-13

[13] You Min-yu, Liu Jia-ming, et al. Embedded feature selection for multi-label classification of music emotions[J]. International Journal of Computational Intelligence Systems, 2012, 5(4): 668-678

[14] Liu Hua-wen, Sun Ji-gui, et al. Feature selection with dynamic mutual information[J]. Pattern Recognition, 2009, 42(7): 1330-1339

[15] Doquire G, Verleysen M. Feature selection for multi-label classification problems[M]. Advances in Computational Intelligence, 2011: 9-16

[16] Cover T M, Thomas J A. Elements of information theory [M]. John Wiley & Sons, 2012

[17] McGill W J. Multivariate information transmission [J]. Psychometrika, 1954, 19(2): 97-116

[18] Brown G. A new perspective for information theoretic feature selection[C]// International Conference on Artificial Intelligence and Statistics. 2009: 49-56

[19] Chen Wei-zhu, Yan Jun, et al. Document transformation for multi-label feature selection in text categorization[C]// Seventh IEEE International Conference on IEEE. 2007:451-456

[20] Zhang Min-ling. ML-RBF:RBF neural networks for multi-label learning[J]. Neural Processing Letters,2009,29(2):61-74

[21] Zhang Min-ling, Zhou Zhi-hua. Multi-label learning by instance differentiation[C]// AAAI. 2007,7:669-674

[22] Zhang Min-ling, Zhou Zhi-hua. ML-kNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition,2007,40(7):2038-2048

(上接第 21 页)

4.2 实验分析

在对定向凝固过程的模拟中,模拟对象的复杂程度决定了总的计算量。对于不同的计算量,GPU 的加速效果会有所差异。为了比较 GPU 相对于 CPU 在定向凝固相场模拟中的优势,本文分别从演化次数和问题规模[1)]方面具体说明,其中 t_{CPU}/t_{GPU} 称为 GPU 相对于 CPU 的加速比,简称加速比。

首先,考察演化次数对加速比的影响。如图 5 所示,当问题规模固定为 $Nx\times Ny=(250\times1500)$时,随着演化次数的增加,加速比逐渐增加。当演化次数从 100 步逐步增加到 100 万步时,加速比逐步稳定在 36.5 左右;当演化次数大于 20 万步后,演化次数对加速比的影响逐步减弱。

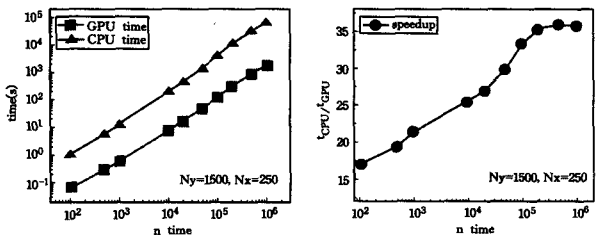

图 5 加速比随演化次数的变化:当问题规模固定时,随着演化次数的增加,加速比逐步稳定在 36.5 左右

其次,考察问题规模对加速比的影响。如图 6 所示,当演化次数固定为 10 万时,随着问题规模从(250×1500)逐步增加到(5000×1500),加速比逐步稳定到 36.5 左右;当问题规模大于(2000×1500)时,加速比基本保持不变。

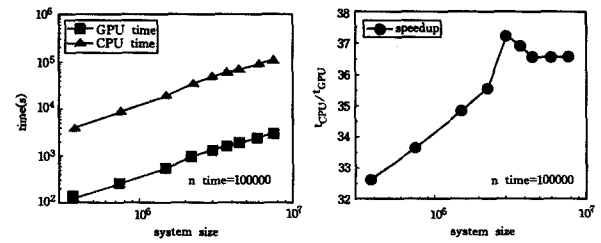

图 6 加速比随问题规模的变化:当演化次数固定时,随着问题规模的增加,加速比逐步稳定在 36.5 左右

通过两组对比实验,可以得出:一方面,随着演化次数和问题规模的增加,GPU 并行计算显著地加快了凝固相场模拟的速度;另一方面,可以明显看出,加速比最终收敛于 36.5 左右。由此可见,只有在大规模计算的时候,GPU 加速的能力才能最大限度地发挥出来。

产生这种现象可能是因为受限于实验的硬件平台。在计算规模较小的时候,GPU 的计算能力部分处于空闲状态,不能完全发挥其优势;随着计算规模的增加,GPU 的能力被充分使用,与 CPU 的串行计算相比,其优势越来越明显;但是,当计算规模达到一定程度时,无论是 GPU 并发执行 kernel 线程块栅格还是 CPU 单线程串行求解,GPU 的硬件资源(主要是 CUDA 核心和共享存储器空间)和 CPU 的硬件资源(主要是 CPU 计算能力)都消耗完全,即硬件的计算能力达到饱和,此时加速比收敛于一个反映了 GPU 和 CPU 固有处理速度的比值。

结束语 本文针对液/固相场模拟过程中,由于大计算尺度和长演化时间导致的模拟计算量巨大的问题,采用 CUDA 计算模型,对相场模拟进行并行化改造,使其适应于 GPU 上的高效并行处理。实验结果表明,在相同的计算尺度和演化时间条件下,GPU 计算与 CPU 计算的加速比可以高达 36 倍。

虽然目前国内类似方法尚处于研究阶段,但本文的研究结果初步表明,GPU 强大的浮点计算能力以及高效的并行处理机制,使得相场理论研究向着大尺度仿真、实时/近实时模拟成为可能;并且,通过采用多核 GPU 的并行协同工作,加速比的理论值还可以进一步提高。

参 考 文 献

[1] Boettinger W J, Warren J A. Simulation of the Cell to Plane Front Transition during Directional Solidification at High Velocity[J]. J. Cryst. Growth,1999,200(3/4):583-591

[2] Wang J C, Zhang Y X, Yang Y J, et al. Phase field modeling for dendritic morphology transition and micro-segregation in multi-component alloys[J]. Science in China Series E: Technological Sciences,2009,52(2):344-351

[3] Lan C W, Shih C J, Lee M H. Quantitative Phase Field Simulation of Deep Cells in Directional Solidification of an Alloy [J]. Acta Mater,2005,53(8):2285-2294

[4] Gurevich S, Karma A, Plapp M, et al. Trivedi. Phase-Field Study of Three-Dimensional Steady-State Growth Shapes in Directional Solidification [J]. Phys. Rev. E. ,2010,81(1)

[5] Wang Y B, Wang Y X, Cheng Z H, et al. Phase-field Simulation of Interface Effect during Grain Nucleation of Solidification Processing [J]. Rare Metal Materials and Engineering, 2012, 41(6):1045-1048

[6] Yamanaka A, Aoko T, Ogawa S, et al. GPU-accelerated phase-field simulation of dendritic solidification in a binary alloy [J]. Journal of Crystal Growth,2010,318(1):40-45

[7] Takaki T, Yamanaka A, Nukada A, et al. Peta-scale phase-field simulation for dendritic solidification on the TSUBAME 2.0 supercomputer [C]// Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis(SC '11). 2011

[8] Glaskowsky P N. NVIDIA's Fermi: The First Complete GPU Computing Architecture[M]// NVIDIA Corporation White Paper. 2009

[9] Losert W, Shi B Q, Cummins H Z. Evolution of Dendritic Patterns during Alloy Solidification: Onset of the Initial Instability [J]. Proc. Nat. Acad. Sci. USA. ,1998,95(2):431-438

1) 为了实现合金定向凝固枝晶列生长过程的模拟加速,本文模拟了 $Nx\times Ny$ 大小的二维剖面,$Nx\times Ny$ 称为问题规模。