第32卷，第1期
2012年1月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 32, No. 1, pp179-182
January, 2012

光谱流量标准化的高效计算

李乡儒

华南师范大学数学科学学院，广东 广州 510631

摘 要 流量标准化是光谱数据挖掘中的一个基本环节，他对挖掘结果的精度和系统的效率均有重要影响，常用方法存在效率较低的问题，为此研究了光谱数据挖掘中流量标准化的算法设计和效率比较问题。首先，探讨了光谱流量标准化技术不同实现方案的渐进效率，给出了实现高效计算的算法，并分析了它们的时间复杂度和空间复杂度。然后，通过SDSS(sloan digital sky survey)的实测光谱数据，横向比较了不同流量标准化算法的效率差异。在光谱流量标准化算法的纵向理论研究中，主要考虑的是计算效率随数据规模增长的变化规律，是在极限意义下进行探讨。在横向实验比较中，考虑重点是不同算法中基本操作时间复杂度的差异及其对算法效率的影响。理论研究和实验结果表明，虽然四种标准化方法 $S_{\max}$，S_{median}，S_{mean} 和 S_{unit} 的渐进效率的类型相同，但对常见的观测规模光谱数据来说，$S_{\max}$ 和 S_{mean} 的效率远远高于 S_{unit} 和 S_{median}，且常用的 S_{unit} 标准化方法效率最低。该研究对于在光谱数据挖掘和开发中，如何根据数据的规模，具体需求，从整体上考虑精度和效率的折衷，以确定合适的流量标准化方法有重要的参考价值。

关键词 光谱数据挖掘；流量标准化；高效计算

中图分类号：TN911.7 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2012)01-0179-04

引 言

随着传感器技术的快速发展，以及2dF，SDSS和LAMOST等大型测谱巡天计划的逐步实施，天文光谱的数据量急速增长，导致了高效天文光谱自动挖掘方法研究的必要性和迫切性[1]。海量天文光谱的数据挖掘是观测天文数据自动处理、信息提取和共享等的关键技术，它在当前数据密集型天文研究中扮演了越来越重要的角色[2,3]。

本工作探讨了光谱数据挖掘中的数据预处理问题。顾名思义，预处理是数据挖掘的准备环节，预处理的质量不仅影响着挖掘的精度/准确性，甚至决定着挖掘系统的稳健性、可用性[4]。因此，预处理是数据挖掘系统的一个关键环节。例如，在天文光谱数据挖掘中，通过预处理环节将光谱数据转换为适合算法需要的格式[3]，去除天光线[5]，矫正或剔除定标畸变[6]，流量标准化[6,7]，纠正或剔除错误数据[3,5,6,8]，去红移并截取公共波段[5,9]。

1 相关研究

Connolly等最早探讨了光谱流量的标准化问题[7]。Connolly提出了三种流量标准化方法，探讨了它们对特征谱和星系光谱分类的影响。结果表明，单位化方法对光谱分类效果最好。它是将光谱看作是高维空间中的一个向量，并通过将光谱向量投影到以原点位圆心的单位超球面上实现流量的标准化。

假设某类天体的理论观测光谱为 x'，由于不同天体在亮度、距离方面的差异，以及观测过程中积分时间的不同，所以实际观测到的光谱 x 往往是理论光谱 x' 的某个倍数。为此，作者提出了如下的流量标准化模型[11]

$$x' = x/\sigma(x) \tag{1}$$

其中，$x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 是观测光谱，x' 是流量标准化后的光谱，$\sigma(x)$ 是标准化因子，它是一个标量。通过定义不同的标准化因子 $\sigma(x)$，可给出各种各样的光谱流量标准化方法。尽管这些方法在理论上是完全等价的，但是，由于实测光谱中噪声的存在，以及计算机存储精度的有限性，导致它们在实际使用中的效果往往不同，有些甚至差别很大。理论分析和光谱分类实验研究结果表明，$S_{\max}$，S_{median}，S_{mean} 和 S_{unit} 四种标准化方法的数值稳定性较好。在这四种方法中，流量标准化因子分别定义如下

$$\sigma_{\max}(x) = \max(x_1,x_2,\cdots,x_n) = x_{(n)} \tag{2}$$

$$\sigma_{\text{median}}(x) = \text{median}(x_1,x_2,\cdots,x_n)$$

收稿日期：2011-05-11，**修订日期**：2011-08-08

基金项目：国家自然科学基金项目(61075033)和广东省自然科学基金项目(S2011010003348)资助

作者简介：李乡儒，1972年生，华南师范大学数学科学学院副教授 e-mail：xiangru.li@gmail.com

$$= \begin{cases} x_{(\lceil n+1\rceil/2)} & n \text{ 为奇数} \\ (x_{(n/2)} + x_{(n/2+1)})/2 & n \text{ 为偶数} \end{cases} \quad (3)$$

$$\sigma_{\text{mean}}(x) = \sum_{i=1}^{n} x_i \Big/ n \quad (4)$$

$$\sigma_{\text{unit}}(x) = \sqrt{\sum_{i=1}^{n} x_i^2} \quad (5)$$

上述 S_{unit} 即是文献[8]中的单位化流量标准化方法，其中 $\tilde{x} = (x_{(1)}, x_{(2)}, \cdots, x_{(n)})^{\text{T}}$ 是通过对 $x = (x_1, x_2, \cdots, x_n)^{\text{T}}$ 的各个分量从小到大排序后得到的向量。

2 光谱流量标准化的高效计算

已有的文献研究中，关于光谱流量标准化的研究主要集中于不同方法在光谱数据挖掘精度/准确性方面的差异及其原因的探讨。在海量光谱数据的自动处理中，除了需要考虑精度问题外，效率也是一个决定算法适用性的关键因素。本文将研究上述四种流量标准化方法 S_{max}，S_{median}，S_{mean} 和 S_{unit} 的实现方案，及其效率差异。在以下探讨中，以 $x = (x_1, x_2, \cdots, x_n)^{\text{T}}$ 表示一条原始观测光谱。

对于 S_{max} 和 S_{median}，最直接的实现方法是首先对光谱的各个测量流量排序(假设从小到大排序)

$$\tilde{x} = (x_{(1)}, x_{(2)}, \cdots, x_{(n)})^{\text{T}} \quad (6)$$

然后，分别取序列中的最后一项或中间的项作为标准化因子(如式(2)和式(3)所示)。常用排序算法及其复杂度如表 1 所示，详细探讨请见文献[10,11]。由此可见，基于合并排序，快速排序和堆排序的计算方法较好。

Table 1 The performance of common sorting methods

排序算法	时间复杂度	空间复杂度
插入排序	$\Theta(n^2)$	$\Theta(1)$
冒泡排序	$\Theta(n^2)$	$\Theta(1)$
选择排序	$\Theta(n^2)$	$\Theta(1)$
合并排序	$\Theta(n\log_2 n)$	$\Theta(n)$
快速排序	$\Theta(n\log_2 n)$	$O(\log n)$
堆排序	$O(n\log_2 n)$	$\Theta(1)$

实际上，在 $\sigma_{\text{max}}(x)$ 和 $\sigma_{\text{median}}(x)$ 的计算中，并不需要对所有的流量分量排序，例如，如果要计算 $\sigma_{\text{max}}(x)$，只需要从 x_1 至 x_n 逐个检查一遍即可，具体算法请见表 2，该算法的时间复杂度和空间复杂度分别是 $\Theta(n)$ 和 $\Theta(1)$。对于 $\sigma_{\text{median}}(x)$，我们的方案是首先采用分区[12]的方法快速排除干扰项，具体算法如表 3 所示，其时间复杂度和空间复杂度亦分别是 $\Theta(n)$ 和 $\Theta(1)$。

Table 2 The algorithm to compute normalization factor σ_{max}

```
σmax=x1;
for i=2:n
    if xi>σmax
        σmax=xi;
    end
end
```

Table 3 The algorithm to compute normalization factor σ_{median}

```
x: 待排序的向量
n: 向量 x 的维数
k=[(n+1)/2];
l=1, r=n, rr=r, p=-1;
如果 x(r)小于 x(l)则对它们互换
i=l,j=r;
while(p~=k)
    while(i<j)
        i=i+1,j=j-1;
        while(x(i)<x(l))
            i=i+1;
        end
        while(x(j)>x(l))
            j=j-1;
        end
        互换 x(i)和 x(j)
    end
    %Cancel the last swap
    if (i>j)
        交换 x(i)和 x(j)
    end

    交换 x(j)和 x(l)
    p=j;
    如果 p>k 则令 rr=p,r=p-1, i=l,
    j=r;否则,如果 p<k,则令
    l=p+1,i=l,j=r;
    如果 x(r)小于 x(l)则对它们互换
end
如果 n 是偶数则按以下步骤计算中值
    min=x(p+1);
    for  i=(p+2):rr
      如果 min>x(i), 令 min=x(i);
    end
    median=(x(p)+min)/2;
否则中值为 median=x(p)
```

Table 4 The algorithm to compute normalization factor σ_{mean}

```
σmean=x1;
for  i=2:n
        σmean=σmean+xi;
end
σmean=σmean/n
```

Table 5 The algorithm to compute normalization factor σ_{unit}

```
σunit=x1^2;
for  i=2:n
        σunit=σunit+xi^2;
end
σunit=σunit的平方根
```

流量标准化因子 $\sigma_{\text{mean}}(x)$ 和 $\sigma_{\text{unit}}(x)$ 的计算不需要特殊技巧，分别按照式(4)和(5)进行即可，实现算法如表 4 和表 5，它们的时间复杂度和空间复杂度均为 $\Theta(n)$ 和 $\Theta(1)$。

3　实验探讨与结论

在上一部分我们研究了流量标准化方式 S_{max}，S_{median}，S_{mean} 和 S_{unit} 的高效计算问题，采用的是纵向、理论探讨方式，所给表 2—表 5 中算法的渐进效率如表 6 所示。下面我们将以横向、实验的方式比较这四种流量标准法方式的计算效率。实验数据采用的是 Sloan 发布的 5 071 条 Galaxy 光谱[13]，所在天区编号是 0267—0276，每条光谱均为一个 3 791维的向量。

Table 6　The performance of the algorithms proposed by this work

标准化方法	时间复杂度	空间复杂度
S_{max}	$\Theta(n)$	$\Theta(1)$
S_{median}	$\Theta(n)$	$\Theta(1)$
S_{mean}	$\Theta(n)$	$\Theta(1)$
S_{unit}	$\Theta(n)$	$\Theta(1)$

在 S_{max}，S_{median}，S_{mean} 和 S_{unit} 计算效率的纵向渐进效率研究中，我们考虑的是计算效率随数据规模增长时的极限增长规律。在这里的横向比较中，我们进一步考虑了不同算法中基本操作的差异，例如，在表 2 的 S_{max} 标准化因子计算中，

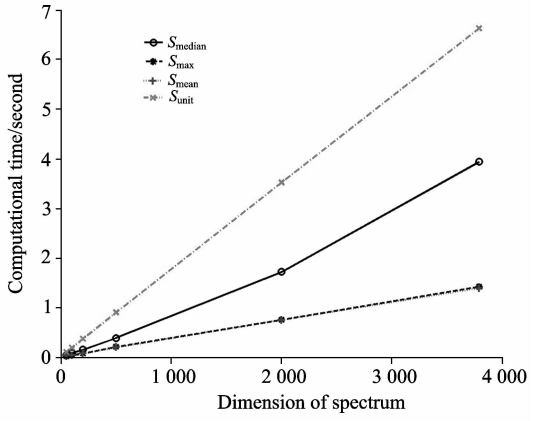

Fig. 1　The computational efficiency of the algorithms in Tables 2—5

基本操作是逻辑比较，而表 5 的 S_{unit} 标准化因子计算中基本操作是求平方，该基本操作的时间复杂度比逻辑判断要小很多。S_{max}，S_{median}，S_{mean} 和 S_{unit} 的横向效率比较结果如图 1 所示。由此可见，S_{max} 和 S_{mean} 的效率远远好于其他两个标准化方法，S_{unit} 的效率最低。所以，在实际光谱数据挖掘中，我们需要根据数据的规模，以及具体问题的需求，从整体上考虑精度和效率平衡，以确定合适的流量标准化方法。

References

[1] YAN Tai-sheng, ZHANG Yan-xia, ZHAO Yong-heng, et al(严太生，张彦霞，赵永恒，等). Progress in Stronomy(天文学进展), 2010, 28(2): 112.

[2] Tsalmantza P, Kontizas M, Rocca-Volmerange B, et al. Astronomy & Astrophysics, 2009, 504(3): 1071.

[3] Ball N M, Brunner R J. International Journal of Modern Physics D (IJMPD), 2010, 19(7): 1049.

[4] Jain A K, Duin R P W, Mao Jianchang. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4.

[5] Yip C W, Connolly A J, Vanden Berk D E, et al. The Astronomical Journal, 2004, 128(6): 2603.

[6] Richards J W, Freeman P E, Lee A B, et al. The Astrophysical Journal, 2009, 691: 32.

[7] Connolly A J, Szalay A S, Bershady M A, et al. The Astronomical Journal, 1995, 110(3): 1071.

[8] McGurk R C, Kimball A E, et al. The Astronomical Journal, 2010, 139(3): 1261.

[9] Vanderplas J, Connolly A. The Astronomical Journal, 2009, 138(5): 1365.

[10] ZHANG Jing(张　静). Journal of Hexi University(河西学院学报), 2010, 26(2): 69.

[11] ZHENG Zong-han, ZHENG Xiao-ming(郑宗汉，郑晓明). The Design and Analysis of Algorithms(算法设计与分析). Beijing: Tsinghua University Press(北京：清华大学出版社), 2005.

[12] Levitin A. Introduction to the Design and Analysis of Algorithms, 2nd Editin. New Jersey: Addison Wesley/Pearson, 2007.

[13] Abazajian K N, Adelman-McCarthy J K, Agüeros M A. The Astrophysical Journal Supplement, 2009, 182(2): 543.

Efficient Computation of Spectral Flux Normalization

LI Xiang-ru

School of Mathematical Sciences, South China Normal University, Guangzhou　510631, China

Abstract　Flux normalization is a key procedure in spectral data mining, and is important for the efficiency and accuracy of automatic processing of massive astronomical spectral data, information extraction and sharing. Since the usual implementation of flux normalizing methods is inefficient, the present work focuses on the algorithm designing of spectral flux normalization. Firstly, the authors investigated the limit efficiency characteristics of the available flux normalization methods, introduced four effi-

cient flux normalizing algorithms, and studied their time complexity and space complexity. Secondly, the authors evaluated the efficiency of the proposed algorithms experimentally and horizontally based on the SDSS (Sloan Digital Sky Survey) released spectral data. In the theoretical research, the main consideration is the computational complexity characteristics of the flux normalization methods when the data size increases unlimitedly. The experimental research focuses on the difference in the computational burden between the basic operations in different flux normalization methods. It is shown that, although the four flux normalization methods S_{max}, S_{median}, S_{mean} and S_{unit} belong to the same limit efficiency type, on the spectra with usual observing scale, S_{max} and S_{median} are much more efficient than S_{mean} and S_{unit}, and S_{unit} is the most inefficient one. This work is helpful for choosing the appropriate flux normalization method based on the size of spectra database and the scientific needs in automatic spectra analysis.

Keywords Spectral data mining; Flux normalization; Efficient computation

(Received May 11, 2011; accepted Aug. 8, 2011)