DOI:10.16016/j.1000-5404.2002.02.035
第24卷第2期 第三军医大学学报 Vol. 24, No. 2
194 2002年2月 ACTA ACADEMIAE MEDICINAE MILITARIS TERTIAE Feb. 2002

论著

文章编号:1000-5404(2002)02-0194-03

基于实验均方误差 *F* 值的基因表达数据最佳聚类评判方法研究

刘月明,易 东,王文昌,张 蔚,颜 杰 (第三军医大学高原军事医学系卫生统计学教研室,重庆 400038)

提 要:目的 近年来产生了一些用于分析基因表达数据的聚类算法,却很少有关于评价聚类算法方法的研究。本研究的目的是尝试建立一个定量地评价基因表达数据聚类结果的方法。方法 本研究提供了一个系统的评价聚类结果的方法,利用我们提出的实验均方误差 *F* 值对几个常见的聚类算法进行比较。结果 利用 *F* 值对类质量的评价和利用已有的生物学知识对类进行分析的结果一致。结论 实验均方误差 *F* 值可以定量地评判用于基因表达数据的聚类算法。

关键词:基因表达数据;聚类算法;均方误差 *F* 值;评价
中图法分类号:R195.1;R394-33 文献标识码:A

Assessment of optimial clustering algofithm for gene expression data based on conditional root-mean-square *F* value

LIU Yue-ming, YI Dong, WANG Wen-chang, ZHANG Wei, YAN Jie (Health Statistics Department of Third Military Medical University, Chongqing, 400038, China)

Abstract: Objective To provide a systematic framework for assessing the results of clustering algorithms for qene expression data *via* conditional root-mean-square (RMS) *F* value and adjust *F* value (a*F*). Methods Partional (CAST, K-averaqe value and iterative) hierarchical single-link, average-link, and complete-link and random Clustering algorithms were applied on 2 established gene expression data sets, Barrnett esophagus and rat CNS data set respectively. Their *F* values and a*F* values were calculated and compained. An algorithm was considered not suitable if its *F* was consistant with that of random cdusteing algorithm. Results This assessments was successfully carried out, on the quality of 7 methods. The results of our assessment were identical to the results based on biological analyses. Conduscion *F* valuse of RMS is able to quantitatively assess the clustering algorithms for gene expression data. We found our quantitative measures of cluster quality to be positively correlated with external standards of cluster quality.

Key words: gene expression data; clustering algorithm; *F* value of RMS; assess

cDNA 和寡核苷酸微矩阵(DNA Microarray)技术的到来导致了生物学研究热点的移位,研究的瓶颈问题已经从数据产生转移到了数据分析方面[1,2]。聚类分析是一种探索性的数据分析方法。根据研究对象(个体)的属性数值特征,采用数学方法对之进行分类整理;再对同一类内个体的共性及不同类间个体的差异作进一步的归纳,从而得到新的规律。评价聚类结果和解释所发现的类与生成这些类同样重要[3]。在关于基因表达数据聚类分析的文献中,是由已知的生物学知识来评价算法是否成功[4]。这样的评判不但效率低,而且对于生物学知识未知的探索性数据束手无策,本研究的目的是建立一套不依赖于先验知识的评判方法。

1 *F* 值与调整 *F* 值的定义与计算

1.1 *F* 值的定义

一个基因表达数据集一般包含 *n* 个被测基因在 *m* 个实验条件下的表达水平。假设一个聚类算法对含有条件数据进行聚类,留出条件 *e* 用来估计算法的预测能力。对其它实验聚类的同时,实际上也对实验 *e* 进行了聚类。假设 *e* 被聚成 *k* 类:*c*1, *c*2, ……, *ck*。定义 *F* 值如下:

$$F(e,k)=\sqrt{\frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i}(R(x,e)-\mu_{Ci}(e))^2}$$

这里 $R(x,e)$ 表示原始数据矩阵里基因 *x* 在实验条件 *e* 下的表达水平。$\mu_{Ci}(e)$ 表示 C_i 类里的基因在条件 *e* 下的平均表达水平。

由于所有 *m* 个条件都可以用做保留在外面的实验条件,这就产生了 *m* 个 *F* 值,*F* 值的总和 $F(\kappa)=\sum_{e=1}^{m}F(e,\kappa)$,是对把数据集聚成 *k* 类时的质量估计。$F(\kappa)$ 越小意味着类内间的变异越小,也就意味着类的质量较高。但随着类的数目增加,每个算法的 *F* 值都呈下降趋势。导致这种情况有两个原因:第一,当对较大的组进行细分时,理论上来说算法能发现更高质量的类;第二,单纯增加类的数目有减小 *F* 值的趋势。因此需要对 *F* 值进行调整,对第二个原因进行纠正。

1.2 调整 *F* 值(Adjust *F* value, a*F*)的计算

假定一个理想模型:*n* 个基因落入 *c* 个真实的种类中,第 *i* 种类包含 $\alpha_i n$ 个基因,这里 $0<\alpha_i<1$,且 $\sum_{i=1}^{m}\alpha_i=1$。第 *i* 种类里的基因在条件 *e* 下的表达水平以 $\mu_{i,e}$ 均数和方差 σ_i^2, *e* 呈独立的正态分布。

当用一个聚类算法把 *n* 个基因聚成 *k* 类($\kappa\geqslant c$)时,如果聚类算法是准确的,那么每类所包含的基因应该来自同一种类,假设这里 $\alpha_i\kappa$ 有个类包含第 *i* 种类的基因,那么 $F(e;\kappa)$ 的期望总和 $F(\kappa)$ 就等于 $\sqrt{\frac{n-k}{n}}\bar{\sigma}$,这里 $\bar{\sigma}=\sum_{e=1}^{m}\sqrt{\sum_{i=1}^{c}\alpha_i\sigma_{i,e}^2}$ 是 σ_{ke} 的加权平均值。

如果上述假设能够满足,且忽略取整误差($\alpha_i\kappa$ 被假设成整数),*F* 值随着类数 *k* 的增长的下降率应该是 $\sqrt{\frac{n-\kappa}{n}}$。据此我

作者简介:刘月明(1974-),男,河南省信阳市人,硕士研究生,主要从事聚类分析、图像处理方面的研究。电话:(023)68752343

收稿日期:2001-03-13;修回日期:2001-06-15

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

们定义 k 类的调整 F 值为 $F(e,\kappa)/\sqrt{\frac{n-\kappa}{n}}$。下面的结果显示的都是调整 F 值。

2　应用

本研究运行了分割算法: Ben-Dor 和 Yakhini 的 CAST 算法[5]、Jain 和 Dubes 的 k-均值算法[3]; 分层算法: 单连接(Single-link), 平均连接(Average-link)和完全连接(Complete-link)算法。此外, 还运行了随机聚类作为评价聚类算法表现的基准。随机聚类就是对基因进行随机分组的过程。如果一个算法的 F 值和随机聚类的 F 值相似, 可以认为该算法的聚类质量很差。

上述的算法是在下面 2 个数据集上运行的: Barrett[6] 的食道数据集(简记 BE)、大鼠 CNS(中枢神经系统)数据集[2]。结果如下。

2.1　BE 数据集结果

图 1 显示的是聚类算法在 BE 数据集上的调整 F 值。当类的数目超过 30 时, 调整 F 值下降的非常缓慢, 所以这里只显示了 1 到 30 个类的情况。在 10 类之前 F 值急剧下降, 这暗示把数据集分成 10 类是有意义的。并且 CAST 和经过平均连接初始化的 κ-均值算法 10 类附近时最接近, 且这时它们的调整 F 值是算法中最好的。

经过上述分析, 把 CAST 算法、平均连接算法和经过平均连接初始化的 k-均值算法应用到整个数据集(所有 10 个条件), 把它分成 10 类。然后利用已有的生物学知识比较这些类。由 CAST 算法和经过平均连接初始化的 k-均值算法正确地把细胞角蛋白聚类到它们相应的类。而在 10 类时 F 值较高的平均连接算法产生了两个只含 2 个和 3 个成员的小类, 并且不能正确把细胞角蛋白聚到相应的类中。因此, 这显示了 CAST 和经过平均连接初始化的 k-均值算法产生了比平均连接更稳健的类, 同时也验证了 F 值分析。

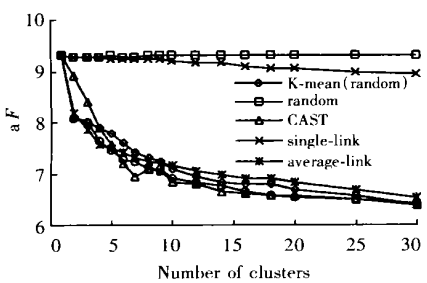


图 1　聚类算法在 **BE** 数据集(**795** 个基因)上的调整 F 值

Fig 1　aF of clustering algorithms on BE data set (795 genes)

2.2　大鼠 CNS 数据集结果

图 2 和图 3 显示了 5 个算法在 1 到 15 类之间的调整 F 值, Wen 等[2] 1998 利用生物学知识把大鼠 CNS 上的基因分类到 4 个家族中, 因此 1 到 15 类是有意义的范围。在图 3 中, 完全连接和平均连接分层算法的调整 F 值比单连接算法小。因为单连接算法倾向于产生"链式"的类, 在这样的类里, 类两端的基因差别可能非常大[7]。该结果和通常的观点相同, 完全连接和平均连接算法要比单连接算法可靠。平均连接聚类比完全连接聚类更能容忍局外值, 这可能是完全连接算法在这个数据集上的 F 值比较小的原因之一。

对聚类算法在把数据聚为 4 类时的结果和 Wen[2] 的分类结果进行比较, 发现经过初始化的 k-均值法及 CAST 法的聚类结果和 Wen 的生物学分类基本一致。而单连接算法产生的结果和 Wen 的结果有较大出入。另外, 对平均连接算法聚类的结果再进行 k-均值聚类显著地改善了原来的结果。这些结果和图中显示的 F 值是一致的, 即 F 值较小的聚类算法产生的类的质量较高。

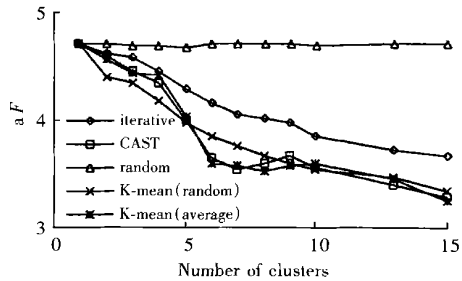


图 2　分割算法在大鼠 **CNS** 数据集(**112** 基因)上的调整 F 值

Fig 2　aF of partitional clustering algorithms on the rat CNS data set (112 genes)

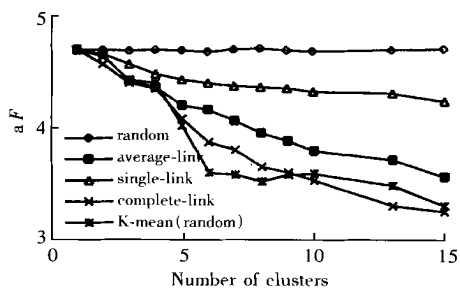


图 3　分层聚类算法在大鼠 **CNS** 数据集(**112** 基因)上的调整 F 值

Fig 3　aF of hierarchical clustering algorithms on the rat CNS data set (112 genes)

3　讨论

基因芯片技术是近年来生物学研究的热点。应用该技术可以在一张芯片上同时进行成千上万个实验, 已经产生了总量巨大的实验数据。如何对这些数据进行准确、有效的分析已成为利用该技术的一个瓶颈问题。聚类是有效的分析手段之一, 且是后续研究的基础。现在已有一些用于基因表达数据分析的聚类算法, 但对聚类结果评判方法的研究很少, 且大多是基于已有的生物学知识对结果进行评判, 不但效率低, 且并非所有的数据都有相应的生物学知识可资利用, 因此对于一个生物学家来说评价一个聚类结果的可靠性并不容易。本研究提供了一个简单的、直接基于数据的方法—F 值评判法, 该法允许在任何数据集比较聚类算法。第 3 节的结果既证实了该法的有效性, 同时也对几个算法在两个数据集上的表现进行了评判: 验证了一般认为平均连接算法和完全连接算法优于单连接

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

算法的观点；还显示了CAST算法在基因表达数据集上有较好表现；运行平均连接算法后重复 k-均值步骤可以提高聚类质量。

如果为了获得不相关的基因或计算配对相似性的话，本研究关于 F 值的定义是用来权衡类内的变异，显然是不适合上述情况的，因此还需要进一步的研究。另外通过导出 F 统计量的概率分布，从而从概率水准评判聚类算法可能是个很有意义的方向。

参考文献：

[1] DeRisi J L, Iyer V R, Brown P O. Exploring the metabolic and genetic control of gene expression on a genomic scale[J]. Science, 1997, 278(5 338): 680—686.

[2] Wen X, Fuhrman S, Michaels G S, *et al*. Large-scale temporal gene expression mapping of central nervous system development[J]. Proceedings National Academy Science USA, 1998, 95(1): 334—339.

[3] Jain A K, Dubes R C. Algorithms for Clustering Data[M]. New Jersey: Wiley, 1988. 1—46.

[4] Eisen M B, Spellman P T, Brown P O, *et al*. Cluster analysis and display of genome-wide expression patterns[J]. Proceedings National Academy Science USA, 1998, 95(25): 14 863—14 868.

[5] Ben-Dor A, Yakhini Z. Clustering gene expression patterns[J]. J Compt Biol, 1999, 6(2): 281—297.

[6] Barrett M T, Sanchez C A, Prevol L J, *et al*, Evolution of neoplastic cell lineages in Barrett oesophagus[J]. Nat Genet, 1999, 22(1): 106—109.

[7] Anderberg M R. Cluster analysis forapplications[M]. San Diego, Ca: Academic Press, 1973. 1—264.

（编辑 薛国文）

个案与短篇

文章编号：1000-5404(2002)02-0196-01

超声及X线在肾结核诊断中的价值比较及超声漏诊误诊原因分析

Diagnotic value of ultrasound and X-ray graphy for renal tuberculosis

黄晓玲[1]，左 松[1]，李 馨[2]（第三军医大学：[1]附属新桥医院超声科，重庆 400037；[2]附属大坪医院野战外科研究所超声科，重庆 400042）

本研究者在进一步探讨超声及X线在肾结核中的诊断价值，并对超声漏、误诊原因进行了详细分析。

1 资料与方法

我院经手术病理证实的住院病人58例，并有完整的病历资料，其中男27例，女31例，年龄18～64岁，平均38.8岁。4例为双肾结核，共62个病肾。术前两周均作过超声检查，52例作过X线肾造影或CT检查。所用仪器为ALOKA-SSD630、620超声诊断仪，3.5 MHz凸阵或线阵探头。

2 结果

2.1 B超、X线诊断正确率及漏误诊率

超声诊断正确率为24.19%，漏误诊率为75.81%；54个病肾的X线（包括平片加静脉肾盂造影、穿刺或逆行造影及CT检查）诊断正确率为46.3%，漏误诊率为53.7%，其中平片加静脉肾盂造影41个病肾，诊断正确14个（34.15%）；X线逆行或穿刺造影4个病肾，诊断正确3个（75.00%）；CT检查12个病肾，诊断正确4个（33.33%）。

2.2 B超、X线主要表现

超声检查出的56个病肾的主要超声表现为肾盂/肾盂输尿管积水（41.07%）、“结石”伴积水（21.43%）、肾缩小为一片强回声伴声影（16.07%）、实质内不规则无回声伴强光团（8.93%）、单纯性囊肿（7.14%）、实质内低回声病灶（5.36%）。41个平片加静脉肾盂造影病肾主要表现为肾盂肾盏不显影（58.54%）、功能差（24.39%）、积水（12.20%）、肾内钙化灶（4.88%）。X线穿刺或逆行造影4个病肾，3个表现为肾内不规则造影剂聚集，肾盏不规则，部分杯口消失。CT检查12个病肾，4个表现为大片钙化影。

3 讨论

文献报道超声诊断肾结核的符合率低者55%[4]，高者91.8%[5]，本组符合率为24.19%，明显低于文献报道，超声与X线综合诊断正确率（46.3%）比较，明显低于后者（$P<0.05$），但若分别与各种X线检查法比较，二者间无显著性差异（P 值均大于0.05）。在肾结核的诊断中，超声及X线的诊断正确率均较低。

本组患者中，超声未提示肾结核者高达69.35%，其主要原因有：①肾结核病理改变的多样性。病理改变的多样性导致声像图表现的多样性。坏死可表现为实质内不规则低回声区，空洞可表现为无回声，纤维化引起的输尿管狭窄可致肾积水，钙化可表现为强光团，故易误诊为占位、囊肿、积水及结石等。②对病史的忽视和对常见病诊断的满足。③对侧肾声像图的误导。由于临床肾结核单侧多见，声像图上见双侧肾脏病变时，检查者易排除肾结核的诊断，但有相当部分病例伴有膀胱结核，结核性膀胱挛缩可引起对侧输尿管末端狭窄或闭塞而致肾积水，双肾病变时排除肾结核诊断亦是误诊的原因之一。

本组漏诊的4个病肾中，3例病变轻微，可能是超声漏诊的主要原因。1例可能为检查者不细心所致。

我们认为提高肾结核的超声诊断正确率，降低误诊率最重要的方法是重视病史，不只满足于常见病如结石、积水等的诊

（下转199页）

作者简介：黄晓玲（1965—），女，重庆市人，硕士，主治医师，主要从事腹部超声诊断方面的研究。电话：(023)68894343

收稿日期：2001-06-10；修回日期：2001-08-10

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net