技术与方法 文章编号:0253-3626(2010)05-0718-03

基于小波去噪的基因表达数据聚类集成算法的研究

徐 源,彭 斌,邓 宇,田考聪
(重庆医科大学公共卫生学院卫生统计教研室,重庆 400016)

【摘 要】目的:为基因表达数据的聚类提供方法学参考。方法:提出了一种小波去噪与模糊聚类集成算法结合起来进行聚类的新算法对酵母细胞的基因表达数据进行聚类,并同模糊聚类集成算法进行比较分析。结果:采用 Micro-precision 方法对新算法和模糊聚类集成算法进行评价。结果表明本文提出的算法具有较高的聚类精度,并且聚类精度高于模糊聚类集成算法。结论:利用本文提出的算法对基因表达数据具有较好的聚类效果。

【关键词】小波去噪;模糊聚类;集成学习;基因表达
【中国图书分类法分类号】R318.04 【文献标识码】A 【收稿日期】2009-09-01

Approach of clustering ensemble in gene expression data based on Wavelet De-noising

XU Yuan, et al
(*Department of Health Statistics, College of Public Health, Chongqing Medical University*)

【Abstract】*Objective*: To provide the methodology reference for the gene expression data clustering. *Methods*: This paper proposed a fuzzy clustering ensemble algorithm based on wavelet de-noising.The new method is applied to the yeast cell's gene expression data, and the clustering results were compared with fuzzy ensemble clustering methods. *Results*: Apply micro-precision to evaluate the clustering results. The research indicated that the proposed method was superior to fuzzy ensemble clustering algotuthm. *Conclusion*: The proposed method has better clustering for the gene expression data clustering.

【Key words】Wavelet de-noising; Fuzzy clustering; Ensemble learning; Gene expression

作者介绍:徐 源(1984-),女,硕士,
研究方向:多元统计分析方法在医疗卫生领域中的应用。
通讯作者:田考聪,男,教授,E-mail:tkc5155@126.com。

检测来综合作出结论。

参 考 文 献

[1] Pliskin N H, Capelli-Schellpfeffer M, Law R T, et al.Neuropsychological symptom presentation after electric injury[J].J Trauma, 1998, 44(4):709-715.

[2] Matsumura H, Kobayashi Y, Mann R, et al. Residual myocardial damage following electrical injuries[J]. J Burn Care Rehabil, 1997, 18(4): 299-305.

[3] Block T A, Aarsvold J N, Matthews K L, et al. The 1995 Lindberg Award, Nonthermally mediated muscle injury and necrosis in electrical trauma[J]. J Burn Care Rehabil, 1995, 16(6):581-588.

[4] Acar K, Boz B, kurtulus A, et al. Using of atomic absorption spectrometry for diagnosis of electrical injury (an experiment rat study)[J]. Forensic Sci Int, 2004, 146(s):3-4.

[5] Pan Y, Xu J, Chen S, et al. Use of the lateral Intercostal perforator-based pedicled abdominal flap for upper-limb wounds from severe electrical injury[J]. Ann Plast Surg, 2006, 56:116-121.

[6] Cancio L C, Jimenez-Reyna J E, Barilla DJ, et al. One hundred ninety-five cases of high-voltage electric injury [J].J Burn Care Rehabil, 2000, 26:331-340.

[7] Cochran A, Edelman L S, Saffle J R, et al. Self- reported quality of life after electrical and thermal injury[J].J Burn Care Rehabil, 2004, 25: 61-66.

[8] Kalkan T, Demir M, Ahemed A S, et al. A dynamic study of the thermal components in electrical injury mechanism for better understanding and management of electric trauma: an animal model[J].Burns, 2004, 30(4):334-340.

[9] 郭景元.现代法医学[M].北京:科学出版社,2000.422.

[10] Chilbert M, Prieto T, Sances A J.Finite element modeling of electrode contacts in electrical injury[J]. Ieee, 1989, 1:323-333.

[11] Wolbarsht ML, 著见刘普和, 李书田译.激光在医学和生物学中的应用[M].北京:科学出版社, 1983.78-79.

[12] Xu X H, Xu H Q, Zhu J Z, et al. A preliminary study of skin electrical injury[J]. Forensic Sci Int, 1995, 73:197-202.

[13] 万立华, 马智华, 张佐才.血管电损伤的实验研究[J].中国法医学杂志, 2001, 16(1):19-22.

(责任编辑:冉明会)

随着人类基因组计划的实施和后基因组时代的到来，生物基因数据呈爆炸性增长之势，如何选择合适的方法从这成千上万的基因数据中提取有用的信息，将成为研究者的巨大挑战。聚类分析作为一种有效的工具，已经广泛地应用于市场研究、模式识别、数据分析和图像处理等多个领域，通过聚类分析，不仅有助于探究基因的未知功能，还可以对基因调控途径和调控网络的研究给予启发。

由于基因表达数据具有高属性维、低样本维以及存在噪声等特点，传统的聚类分析可能难以达到理想的聚类效果。集成学习[1](Ensemble learning)是近几年来机器学习领域发展的新趋势，它是将不同的聚类算法或者同一聚类算法采用不同的参数或初始值得到的结果进行集成，得到最终的聚类结果。集成学习在分类和预测上的成功应用，使得越来越多的学者把目光投向如何利用集成学习来提高聚类分析的效能。A Fred 于 2002 年首先对层次聚类算法进行集成，之后，各种集成算法层出不穷，D.Frossyniotis 选取 boosting 算法对聚类进行集成，而 X.Z Fem 使用随机投影法将高维数据投影到低维空间，通过多次投影获得若干个聚类成员，再用 EM 算法对聚类成员进行集成，得到最终的聚类结果等等。实践证明，聚类的集成算法能很好的提高聚类算法的鲁棒性和稳定性，从而得到比单一聚类算法更为优越的结果。本文将集成学习应用到基因表达数据中，对基因进行聚类，得到了良好的聚类效果。

1 基本原理

1.1 基于小波变换的去噪处理

由于基因表达数据在测定过程中，容易引入噪声，而噪声对象不属于任何的簇，在聚类过程中会产生干扰，所以为了反映生物本质的聚类结果，必须进行去噪处理。小波分析[2]是 20 世纪 80 年代末 90 年代初兴起的一种数学理论，不仅为纯粹数学的研究提供了强有力的工具，还被广泛应用于信号分析、图像处理、量子物理、地震勘探以及非线性科学等领域，并在信号滤波、去噪、压缩等方面表现出众。小波变换是一种时－频域变换，能有效地从原始信号中提取信息，并通过伸缩和平移等运算对信号进行多尺度分析。由于它可对原始信号的高频部分在时域或频域上采用逐渐精细的取样步长进行分析，从而可以聚焦到分析对象的任意细节，并在消除噪声的同时，并未光滑掉数据的主要特征，故特别适用于滤噪。

小波去噪的基本思想是通过小波变换，把信号分解为低频信号和高频信号，低频成分给出了信号的特征，高频成分则与噪音和扰动连在一起。将信号的高频成分去掉，信号的基本特征仍然可以保留，这样就剔除了数据中的噪声。

一个含噪的一维信号模型可表示为如下形式：

$S(k)=f(k)+\varepsilon e(k),k=0,1,\cdots,n-1$

其中，s(k)为含噪信号，f(k)为有用信号，e(k)为噪声信号。e(k)是一个 1 级高斯白噪声，通常表现为高频信号；f(k)通常为低频信号或者是一些比较平稳的信号。

1.2 FCM 集成

1.2.1 聚类集成 聚类集成的具体表达如下：若有 n 个数据点 $x=\{x_1,x_2,x_3,\cdots,xn\}$，对数据集 x 用 H 次聚类算法得到 H 个聚类结果(也叫聚类成员)，$H=\{\pi_1,\pi_2,\pi_3,\cdots,\pi_n\}$，其中 $\pi_i=(i=1,2,3,\cdots,H)$为对第 i 次算法得到的聚类结果。设计一种共识函数 Γ，对这 H 个聚类成员的聚类结果进行合并，得到一个最终的聚类结果 π'。如图 1 所示

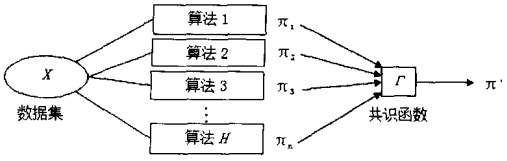


图 1 聚类集成示意图

由图 1 可以看出，聚类集成的关键就是生成有差异的聚类成员以及选择合适的共识函数。

1.2.1.1 产生聚类成员 聚类成员可以通过不同的聚类算法来产生也可以使用同一聚类算法选择不同的初始值或参数来产生，但前者由于聚类原理不同，产生的聚类成员形式会不同，就会给共识函数的设计带来较高的复杂性，一般不予考虑。本文结合基因表达数据中基因多面性的特点，选取模糊 c 均值算法[3](Fuzzy c-means，FCM)来进行聚类，FCM 算法是在基于目标函数的聚类算法中理论最完善、应用最广泛的算法，该算法是从硬 c 均值(Hard c-means，HCM)聚类算法中发展而来的。FCM 算法能自动对数据样本进行分类，通过优化模糊目标函数得到每个样本点相对类中心的隶属度，从而决定样本点的分类。

1.2.1.2 构建共识函数 为了将聚类成员集成为最终的聚类结果，需要对产生的聚类成员进行集成。本文选取基于权值的选择投票法[4]作为集成的共识函数。投票法因其原理简单、易于应用而广泛应用于分类中。因聚类是无监督学习，经过多次聚类后得到的簇标记是不确定的，如同一簇的数据在一个聚类中被置于簇标记 1，而在另一个聚类中，可能被置于簇标记 2，所以在进行投票之前必须先统一标记向量。

具体表述如下：

①从产生的聚类结果中随机选择一个标记向量作为基准，然后将其他标记向量与选择的基准标记向量进行统一。

②将相应的标记向量进行转换，得到统一的新的标记向量。

③对每一个模糊聚类的结果进行投票，以个体标记向量的权值作为选择聚类集成个体标记向量的依据。当某个标记向量的权值低于某个预设阈值时，则该个体标记向量将不参与最后的集成。

④将挑选出的个体标记向量再基于权值进行投票集成，也就是看对于某个数据点，如果聚类集成中有大多数的聚类都把它归为一类，则最终就归为哪一类，从而得到最终的聚类结果。

2 结果与评价

2.1 数据来源

选取酵母全基因组在有丝分裂周期中的 17 个时间点的基因表达数据，该数据由 6 000 多条酵母细胞中起调控作用的基因表达数据组成，本文用 Cho 从这些数据中提取的 232 组数据[5]进行聚类。这 232 组数据根据不同的时相可分为 4 类：G1（116 条），S（49 条），G2（33 条），M（34 条）。

2.2 算法的实现

2.2.1 小波去噪 由于小波分析所用的函数具有多样性，因此在实际应用过程中，选择最优的小波基是至关重要的步骤，常用的小波有：Haar 小波、Daubechies（dbN）小波、Mexican hat 小波、Motlet 小波等等，目前主要是通过比较小波分析的处理结果和理论结果的误差来评价小波基的好坏。本文通过若干次实验的比较，选用 daubechies（dbN）（N=3）作为小波母函数进行 2 层小波分解，得到的低频分量既可以表现原始信号的大致概貌，又对于峰值有很好的体现。小波去噪能进行并行处理，这里选择了其中 1 个基因的小波去噪图，如下：

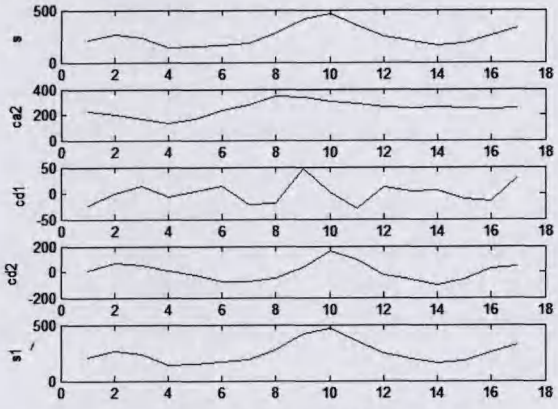


图 2 小波变换去噪图

图 2 中，s 为原始信号，即为 17 个时间点的酵母基因表达数据，ca2 表示进行 2 层小波分解后得到的低频信号，该信号就是原信号中的有用的部分，反映原信号的幅值和基本变化趋势；cd1 和 cd2 为分解后得到的高频信号。S1 为重构信号。由图 1 可知：重构信号与原始信号在变化趋势上基本保持一致，在峰值处略显光滑，说明重构信号在去除了噪声的同时较好的保留了信号的原始特征。原信号的分解关系为：s=ca2+cd1+cd2，重构则为分解的逆过程。

2.2.2 FCM 集成 对去噪后的基因表达数据采用 FCM 算法进行聚类分析，鉴于加权指数 m 的最佳取值目前尚缺乏理论指导，故根据查阅大量文献的经验值选取 m=2，通过改变初始聚类中心，根据最近最大划分原则把每个对象分配到隶属度最大的聚类中，由此获得若干个聚类结果，再通过投票法对若干聚类结果进行集成，得到最终的聚类结果。

2.3 评价方法[6]

本文采用 Modha 提出的 Micro-precision 评价方法，该方法利用了数据的内部信息来评价聚类结果的好坏，其度量标准 Micro-precision 定义如下：

$$\text{Micro-precision}=\frac{1}{n}\sum_{i=1}^{k}\alpha_i$$

式中，n 为数据集样本总数；k 为聚类的簇数；α_i 为聚类的簇与已知数据集类别对应后，簇当中被正确归为相应类别的样本个数，当然，Micro-precision 的值越大，表明在该数据集上聚类效果越好。这种评价策略适用于比较产生固定聚类个数的算法性能。因此，本次实验 FCM 的聚类个数设置为已知的类别数 4。

将本文提出的算法与单纯的 FCM 算法集成进行比较，分别构造了规模为 5、10、15、20、30 的集成，每次测试都重复 10 次，取平均的 Micro-precision 值。

从图 3 可以看出，本文提出的算法聚类精度明显高于单纯的 FCM 集成算法，并且随着集成规模的增大，聚类性能也有相应的提高，对比结果显示本文提出的方法是行之有效的。

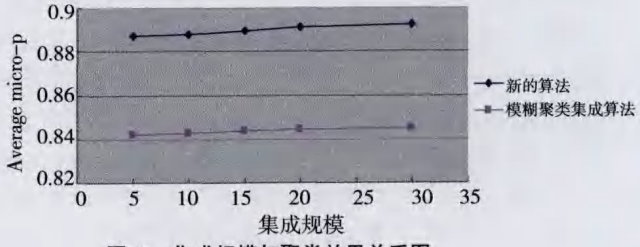


图 3 集成规模与聚类效果关系图

3 讨 论

FCM 在处理较大数据量时具有较理想的扩展性和效率，但容易陷入局部最优，且结果受噪声和异常值的影响较大。为了解决这些难题，本文设计并实现了一种基于小波变换去噪的模糊聚类集成算法，并从一定程度对算法进行了评价。结果表明，该算法具有较高的准确性和稳定性，是一种有效的基因表达数据聚类方法，为进一步分析基因的生物学功能提供参考，但在进行加权投票过程中，如何降低统一标记向量时的时间复杂度是需要改进的地方。

参 考 文 献

[1] A.Strehl，J Ghosh.Cluster ensembles- a knowledge reuse framework for combining multiple partitions[J].Journal of Machine Learning Research，2002，3：583-617.

[2] 飞思科技产品研发中心.小波分析理论与 MATLAB R2007 实现[M].电子工业出版社，2007.354-358.

[3] 李柏年.模糊数学及其应用[M].安徽：合肥工业大学出版社，2007.61-63.

[4] Zhou Z H，Tang W.Clusterer ensemble[J].Knowledge-Based Systems，2006，1(19)：77-83.

[5] Cho R J，Campbell M J.Winzeler E A，et al.A genome-wide transcriptional analyses of the mitotic cell cycle[J].Molecular Cell，1998，2：65-73.

[6] Modha D S，Spangler W S.Feature weighting in k-means clustering[J].Machine Learning，2003，52(3)：217-237.

（责任编辑：关蕴良）