文章编号 1672-6634(2021)06-0020-09 **DOI** 10.19728/j.issn1672-6634.2021.06.0003

基于聚类的多样本复发拷贝数变异检测算法

陈念华,袁细国

(西安电子科技大学 计算机科学与技术学院,陕西 西安 710071)

摘 要 拷贝数变异是人类基因组中一种重要的结构变异类型。不同样本中相同区域出现的拷贝数变异称作复发拷贝数变异。研究表明,复发拷贝数变异与人类复杂疾病紧密关联。提出一种基于聚类思想的多样本复发拷贝数变异的检测算法,该算法首先提取两种与复发拷贝数变异密切相关的特征:即多样本中每个位点的拷贝数变异比率和拷贝数变异幅度均值,然后利用聚类算法在这两种特征上进行聚类,根据聚类结果找出发生复发拷贝数变异的位点。通过两种模拟数据来评估该算法的性能,同时与三种同行方法进行比较,结果表明该算法具有较好的检测性能;本文还将该算法应用至两种真实数据,检测结果中包含一定数量的疾病相关基因,这表明本文所提算法的有效性。

关键词 复发拷贝数变异;聚类算法;多样本;疾病相关基因

中图分类号 N39;TP311

文献标识码 A

开放科学(资源服务)标识码(OSID)

A Recurrent Copy Number Variation Detection Algorithm From Multi-sample Based on Clustering

CHEN Nianhua, YUAN Xiguo

Abstract Copy number variation is an important type of structural variation in the human genome. The copy number variation that occurs in the same region in different samples is called recurrent copy number variation. This paper proposes a cluster-based algorithm to detect recurrent copy number variation from multiple samples. The algorithm first extracts two features that are closely related to recurrent copy number variation: Copy number variation ratio of each probe in multiple samples and the copy number variation amplitude of each probe, then use clustering algorithm to cluster these two features, and find out the probes of recurrent copy number variation based on the clustering results. This paper evaluates the performance of the algorithm through two kinds of simulation data, and compares with three peer methods at the same time. The results show that the algorithm has better detection performance. This paper also applies the algorithm to two kinds of real data, and the detection results contain a number of disease-related genes, which shows the effectiveness of the algorithm proposed in this article.

Key words recurrent copy number variation; clustering algorithm; multiple sample; disease-related genes

收稿日期:2021-03-01

基金项目:国家自然科学基金面上项目(61571341);山东省社会科学规划数字山东研究专项(20CSDJ09)资助

通讯作者:袁细国,男,汉族,博士,副教授,研究方向:生物信息计算,E-mail:xiguoyuan@mail.xidian.edu.cn。

0　引言

拷贝数变异(Copy Number Variation，CNV)是人类基因组中一种重要的结构变异类型，长度通常在 1K base pairs (bp)到 3Mbp 之间，包括拷贝数扩增(amplification)和拷贝数缺失(deletion)两种类型[1,2]。人类基因在正常情况下是二倍体，所以对于人类基因组来说，拷贝数扩增是指基因组区域的拷贝数从正常二倍体到多倍体的变化，拷贝数缺失则是基因组区域中拷贝数减少的变异，若拷贝数缺失至单倍体，称作杂合性缺失；若拷贝数缺失至 0，则称作纯合性缺失.研究表明，CNV 在人类基因组中十分常见，它会引起基因表达发生异常，与人类复杂疾病紧密关联，例如自闭症[3]、精神分裂症[4]、自身免疫性疾病[5]以及癌症[6]等疾病。

自 1975 年第一代 DNA 测序技术开创至今，人类已经积累了大量的测序数据，这使得利用计算机技术对这些数据进行分析成为可能。相比于直接用医学手段检测 CNV，利用计算机技术检测 CNV 更加便捷，成本也十分低廉。当下检测 CNV 的主要难点在于如何区分驱动 CNV[7]和随机 CNV。所谓驱动 CNV，是指对疾病有直接影响或者关联较大的 CNV，找出驱动 CNV 对理解疾病的发病机理有很大帮助；随机 CNV 则是指在基因中随机出现、与疾病的发生关联不大的 CNV。在多样本检测[8]的前提下，CNV 按照在不同样本中发生的频率不同可以分为复发 CNV[9]和个体 CNV[10]，其中复发 CNV 指在不同患者基因组中相同位置发生的 CNV，而个体 CNV 在不同患者基因组中发生的位置则是随机的.研究表明，复发 CNV 更有可能是驱动 CNV，即更有可能包含疾病相关基因，因此本文算法的目标就是从多样本数据中检测出复发 CNV。

当前有许多检测复发 CNV 的方法，例如 PLA(Piecewise-constant and low-rank approximation for identification of recurrent copy number variations)[11]是将多样本 CNV 检测问题转化为矩阵分解问题，其中原始数据矩阵被分解为低秩分量，稀疏分量和噪声分量。这 3 个成分分别对应复发 CNV，个体 CNV 和随机噪声。FLLat(A fused lasso latent feature model for analyzing multi-sample aCGH data)[12]则是使用潜变量模型对基于阵列的比较基因组杂交技术(array-based Comparative genomic hybridization，aCGH)数据进行建模，其中每个样本均通过固定数量特征的加权组合来建模。这些特征代表了样本组 CNV 的关键区域，并与权重相结合，描述了每个单独样本中的 CNV 区域。SAIC(Genome-wide identification of significant aberrations in cancer genome)[13]使用置换检验方法来评估每个位点的重要程度，以此来检测复发 CNV。

如前所述，现有多样本 CNV 检测方法更关注数据的数学特性，而忽略了数据所包含的生物特性，因此本文提出一种基于聚类的从多样本中检测复发 CNV 的新算法 DBSCAN-CNV(A recurrent copy number variation detection algorithm from multi-sample based on clustering)，该算法首先提取两种与复发 CNV 的发生紧密关联的特征，分别是每个位点发生 CNV 样本的比率和每个位点的幅度均值，然后根据这两个特征进行聚类.由于发生复发 CNV 的位点相较于正常位点仅占少数，在整体数据中属于异常点，因此本文采用的聚类方法为 DBSCAN (A density-based algorithm for discovering cluster in large spatial databases with noise)，DBSCAN 的优势在于可以对任意形状的簇进行聚类，并且如果对参数恰当地设定，它可以将噪声点剔除，这可以解决发生复发 CNV 位点在全部数据中占比低的问题。

本文分别将该算法应用在模拟数据和真实数据上，并与三种同行方法进行比较(PLA、FLLat、SAIC)，实验结果表明，本算法对于复发 CNV 的检测性能确实有一定提升。

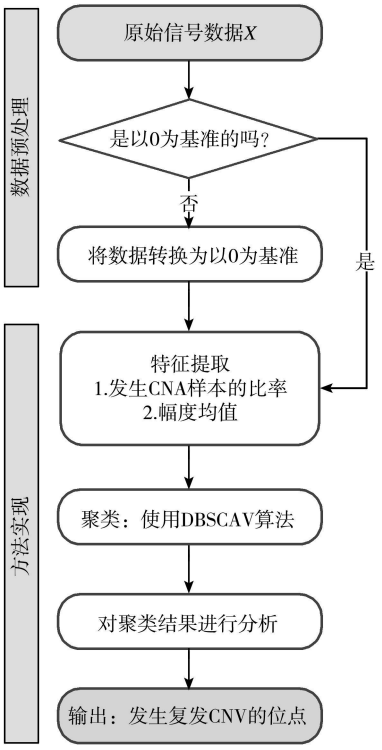

图 1　DBSCAN-CNV 的主要步骤

1　方法

本文算法的流程如图 1 所示，该算法通过以下 4 个主要步骤实现对复发 CNV 的检测：(1) 数据预处理，这一步主要是将数据中的拷贝数信息(以 2 为基准)，转化为以 0 为基准的数据，即将原始数据除以 2

然后取对数,这样方便对 CNV 的类型做区分:信号值小于 0 代表缺失,大于 0 代表扩增(若数据本身就是以 0 为基准的,跳过该步骤);(2) 特征提取;(3) 根据上一步得到的特征进行聚类;(4) 根据聚类结果获得复发 CNV 的检测结果。下面是对第(2)(3)步骤的详细说明。

1.1 特征提取

由于测序错误、正常细胞污染等噪声的存在,原始数据往往呈现为杂乱无章的信号序列,因此本文采用循环二元分割算法(Circular binary segmentation, CBS)[14] 对每个样本进行分段平滑.如图 2 所示,分段平滑后会将原始单个样本数据分为多个连续区域,区域内部的信号值是相同的。

1.1.1 每个位点发生 CNV 的样本比率。在对每个样本进行分段平滑之后,根据分段区域内的信号值越大,则代表该区域的拷贝数越大的原理,选定合适的阈值,判断每个样本在每个位点处是否发生 CNV。对于拷贝数扩增,选定正阈值,分段内信号值若大于该阈值则认为该分段内的所有位点均发生拷贝数扩增;相应的,对于拷贝数缺失,选定负阈值,分段内信号值若小于该阈值则认为该分段内的所有位点均发生拷贝数缺失。

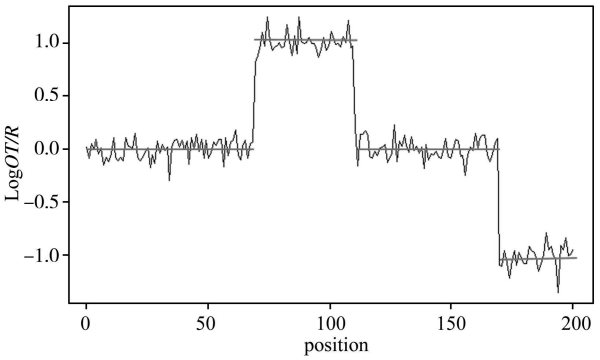

图 2 使用 CBS 对原始数据进行分段平滑,曲线是原始数据,直线是平滑过后的数据

经过以上操作可以得到每个样本在各位点发生 CNV 的情况,据此可以在每个位点计算发生 CNV 的样本占总样本的比率,即

$$freq(i)=count(i)/S\ , \tag{1}$$

其中 $count(i)$ 指在第 i 个位点处发生 CNV 的样本数,S 指总样本数。因为复发 CNV 正是指那些在不同样本间发生频率较高的相同 CNV,因此每个位点发生 CNV 的样本比率是检测复发 CNV 的重要特征。

1.1.2 每个位点的幅度均值。在经过数据预处理之后,数据都是以 0 为基准的(0 代表拷贝数为 2),不论是大于 0 还是小于 0 都代表拷贝数发生了变异.对每个位点处各个样本的信号值取绝对值,然后再取均值,便得到每个位点的幅度均值,其代表了每个位点的拷贝数均值与正常拷贝数偏离的程度,即

$$ampli(i)=\sum_{j=1}^{S}|data[j,i]|/S\ , \tag{2}$$

其中 $data[j,i]$ 表示在第 j 个样本,第 i 个位点处的信号值,S 指总样本数.幅度均值越大,说明该位点的拷贝数偏离正常值越多,因此幅度均值也是检测复发 CNV 的重要特征。

1.2 DBSCAN 聚类

经过上述操作,我们得到每个位点发生 CNV 的比率以及每个位点的幅度均值这两个特征,接下来需要根据这两个特征对所有位点进行聚类。本文采用的聚类算法 DBSCAN[15],是一种基于密度的聚类方法,通过每个点 Eps 半径之内点的个数来衡量每个点的密度,如图 3 所示,可以对任意形状的数据进行检测。

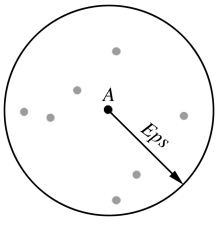

图 3 基于中心的密度,点 A 的密度是 9(包含 A 本身)

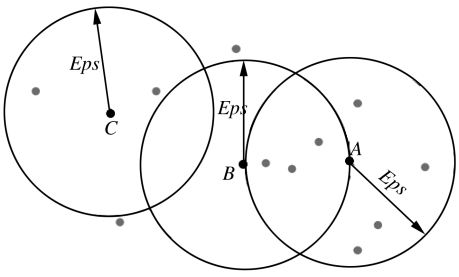

图 4 若 $MinPts=7$,则 A 是核心点,B 是边界点,C 是噪声点

基于密度的聚类方法将数据集内的点分为核心点、边界点和噪声点三类.核心点是在基于密度的簇内部的点.点的邻域由距离函数和距离参数 Eps 决定。如果在一个点的半径为 Eps 的邻域内,包含的点的个数超过阈值 $MinPts$,则这个点是一个核心点;若某个点落在某个核心点的邻域内,并且该点是非核心点,则这个点为边界点;噪声点是既非核心点也非边界点的任何点。图 4 是这三种点的图示。

本文距离函数使用欧氏距离，两点之间的距离由公式定义

$$Dist(i,j)=\sqrt{(ampli[j]-ampli[i])^2+(freq[j]-freq[i])^2}\ 。\tag{3}$$

给定核心点、边界点和噪声点的定义后，DBSCAN 算法可以非形式地描述如下：任意两个相互距离在 Eps 之内的核心点将放在同一个簇内。落在某个核心点邻域内的边界点和该核心点放在同一个簇内.噪声点不属于任何一个簇。下面是 DBSCAN 算法的详细描述：(1) 将所有点标记为核心点、边界点或噪声点；(2) 删除噪声点；(3) 为距离在 Eps 之内的所有核心点之间赋予一条边；(4)每组连通的核心点形成一个簇；(5) 将每个边界点指派到一个与之关联的核心点的簇中。

如前文所述，虽然 DBSCAN 的实现十分简单，但是检测结果十分依赖半径 Eps 的设定。如果设定的半径足够大，则所有点的密度都等于数据集中所有点的个数；类似地，如果半径太小，则所有点的密度都是 1 (仅包含该点本身)。因此，可以通过观察点到它的第 k 个最近邻的距离(称为 k-距离)来选取合适的 Eps。对于属于某个点的簇，如果 k 不大于簇的大小的话，则 k-距离将很小。然而对于不在簇中的点(如噪声点)，k-距离将相对较大。因此，如果我们对于某个 k，计算所有点的 k-距离，以递增次序将它们排序，然后绘制排序后的值，则我们会看到 k-距离的急剧变化，如图 5 所示。

选取 k-距离发生急剧变化的点对应的 k-距离作为 Eps 是一个比较合适的值。如果我们选取该距离为 Eps 参数，而 k 的值作为 $MinPts$ 参数，则 k-距离小于 Eps 的点将被标记为核心点，而其他点将被标记为噪声或边界点.由[15]可知，$k=4$ 对于大多数数据集都是一个合适的参数设定，因此本文算法默认设 $k=4$。Eps 默认取排序后的 k-距离数组中第 $turn$ 个位置的值，$turn$ 定义为

$$turn=\mathrm{P}\cdot turnPercent,\tag{4}$$

其中 P 是 k-距离数组的长度，$turnPercent$ 是 k-距离导数骤增的点与 k-距离数组长度的比值，经过实验默认取 0.9625。

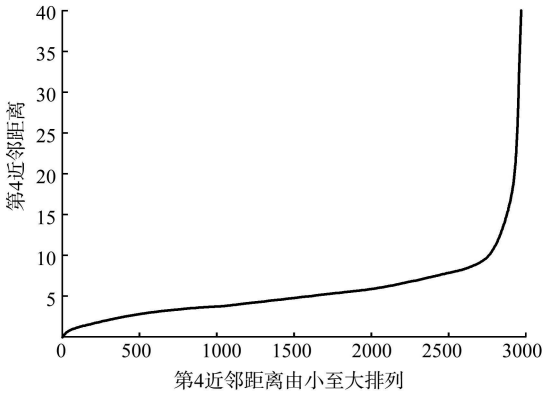

图 5　k 距离的变化趋势

因为发生复发 CNV 的位点在所有位点中所占比率很低，并且其特征与正常位点有显著差异，因此 DBSCAN 的聚类结果中最大的簇代表非复发 CNV 位点，而剩下的簇代表发生在不同位置处的复发 CNV 位点。由于本文的目标是检测出发生复发 CNV 的位点，所以本文将 DBSCAN 聚类结果中除了最大簇以外的簇都作为检测结果，噪声点也视作检测结果，至此对复发 CNV 的检测全部完成。

2　实验结果

为了评估 DBSCAN-CNV 算法对复发 CNV 的检测性能，本文将 DBSCAN-CNV 应用在模拟数据上，并将 PLA、FLLat、SAIC 也应用在这些数据上进行比较.除此之外，本文还将 DBSCAN-CNV 应用在真实数据上，看是否可以检测出已被验证过的疾病相关基因，以此来验证该算法的可用性.接下来是对这些实验的详细说明。

2.1　模拟数据

本文实验分别生成了两种模拟数据，一种是根据文献[16]的描述生成的包含六种场景的高频率复发 CNV 数据，另一种则是本团队设计的相对低频的复发 CNV 数据。

2.1.1　*高频率模拟数据*。在文献[16]里，作者详细地定义了 6 种不同的复发 CNV 场景。本文根据其描述，在每种场景下生成 50 组数据，每组数据是 50×2000 的矩阵，其中 50 代表样本数，2000 代表位点数，即每一行数据都代表一个样本.在生成数据时，将未发生 CNV 的位点的信号值设为 0；复发 CNV 区域位于 750-1250 位点之间，其模式参考图 6。

将扩增区域和缺失区域位点的信号值分别设为 1 和−1。每个样本还需要在非复发 CNV 区域随机选取一个位置，添加一个长度为 200 的个体 CNV，其值从{−2，−1，1，2}中随机选取。最后再向整个数据添加噪声水平为 1 的高斯噪声，图 7 是场景 1 和场景 2 模拟数据的生成过程示例。

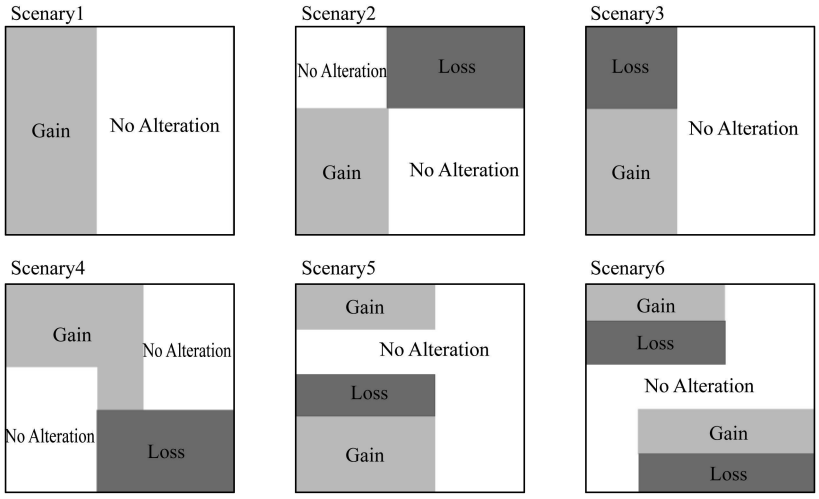

图 6　在 Rueda and Diaz-Uriarte (2010)里定义的六种常见复发 CNV 的模式. 每个场景的纵轴代表样本,横轴代表位点

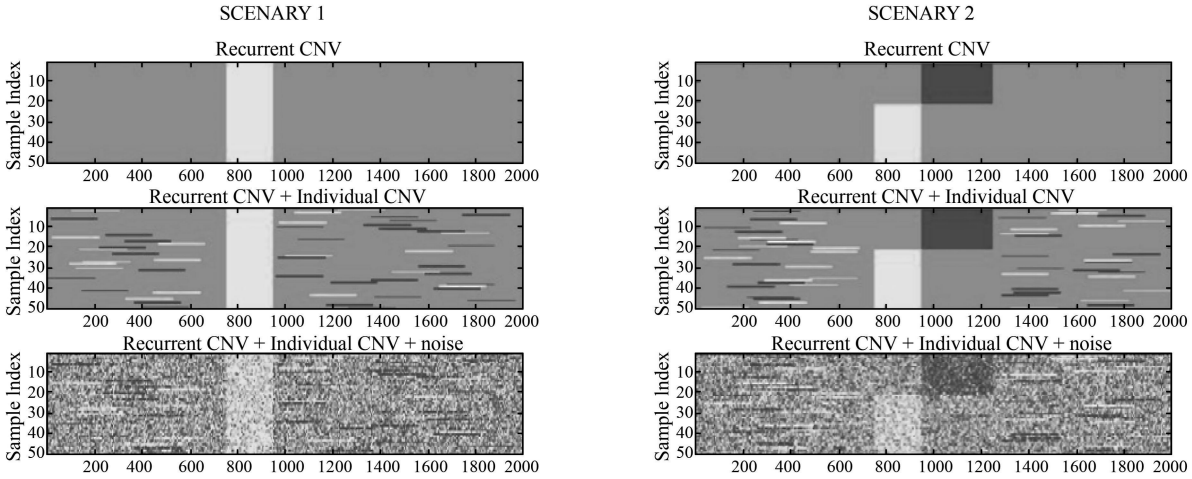

图 7　场景 1-2 模拟数据的生成过程

本文用灵敏度(*sensitivity*)和准确率(*precision*)来评估各个方法对模拟数据的检测性能,其中灵敏度和准确率的含义由公式定义

$$sensitivity = \frac{TP}{TP + FN}, \tag{5}$$

$$precision = \frac{TP}{TP + FP}。 \tag{6}$$

图 8 是四种方法的检测结果图示,横轴为准确率,纵轴为灵敏度,图中曲线是 $F1\text{-}score$ 等高线,$F1\text{-}score$ 是准确率和灵敏度的调和平均值,其定义为

$$F1\text{-}score = \frac{2 \times precision \times sensitivity}{precision + sensitivity}, \tag{7}$$

$F1\text{-}score$ 值越大,说明算法性能越好,对应到图中就是越靠近右上方的点,$F1\text{-}score$ 值越大。

从图中可以看出,除了在场景 4 里 FLLat 和 DBSCAN-CNV 的 $F1\text{-}score$ 相近,在剩下的五种场景里 DBSCAN-CNV 的 $F1\text{-}score$ 值都明显比另外三种方法要大。比如在场景 3 里,虽然 PLA、FLLat 和 DBSCAN-CNV 的灵敏度几乎都达到了 1,但是 PLA 的准确率只有 0.657,FLLat 的准确率是 0.801,而 DBSCAN-CNV 的准确率却达到了 0.98;又比如在复发 CNV 模式比较复杂的场景 6 里,另外三种方法中 $F1\text{-}score$ 最高的 FLLat 也只有 0.810,而 DBSCAN-CNV 的 $F1\text{-}score$ 却达到了 0.908,其中灵敏度是 0.8424,准确率是 0.96。

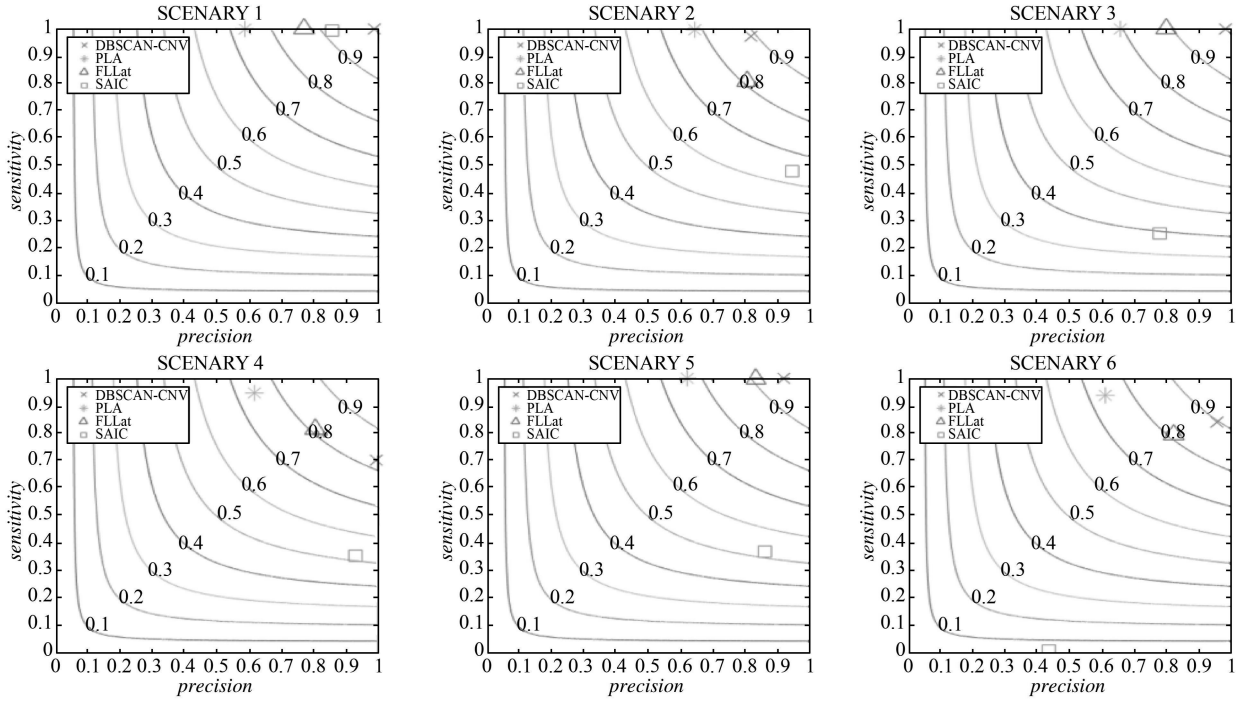

图 8　不同方法对场景 1-6 模拟数据的复发 CNV 检测结果的性能比较

2.1.2　低频率模拟数据。低频率数据按照以下几个步骤生成：(1) 设定复发 CNV 的区域、变异样本所占比率以及拷贝数扩增(或缺失)的大小；(2) 添加个体 CNV 噪声；(3) 添加肿瘤纯度噪声；(4) 添加高斯噪声.接下来是对这些步骤的详细介绍。

在低频率的模拟数据里，本文按照图 9 的模式进行复发 CNV 的模拟。每组数据是 100×2000 的数组，其中 100 代表样本数，2000 代表探针数，即每一行数据代表一个样本。在生成数据时，首先使所有样本的所有位点都为 2(代表正常二倍体)，然后将拷贝数扩增变异区域设定在 100-149、500-529、900-919 位点之间，设定发生 CNV 的样本分别占总样本数的 0.2、0.25、0.20，拷贝数分别扩增到 6、4、5；将拷贝数缺失变异区域设定在 1100-1149、1500-1529、1900-1919 位点之间，设定发生 CNV 的样本分别占总样本数的 0.15、0.2、0.25，拷贝数分别缺失至 0、0/1、1。

图 9　低频率模拟数据的复发 CNV 区间分布

设置复发 CNV 的变异区间后，在每个样本的任意位置添加大小为 100 的个体 CNV，其拷贝数在{0、1、3、4}中任意选取.在真实数据中，通常得到的信号数据并不是全部来自肿瘤细胞的，而是在正常细胞和肿瘤细胞混杂的情况下进行信号的测量，这就导致了信号数据是被正常细胞“污染”过的数据.肿瘤纯度指肿瘤细胞占所有细胞的比率，肿瘤纯度越高，说明数据受到正常细胞的“污染”越少。本文为了模拟这种在真实数据中不可避免的噪声，每个样本数据在 0.3-0.7 之间选择一个肿瘤纯度，然后与正常的拷贝数进行加权平均，这样便得到添加肿瘤纯度噪声之后的数据，最后对每个样本添加指定水平的高斯噪声。

在上述过程中，肿瘤纯度有 0.3-0.7 五种选择，高斯噪声有 0.2 和 0.4 两种选择，通过对这两种参数选择的不同组合，共生成十组数据，每组有 50 个 100×2000 的模拟数据。

为了可量化地比较 DBSCAN-CNV、PLA、FLLat、SAIC 这四种方法在这十组数据上的检测性能，这里依旧使用灵敏度(*sensitivity*)和准确率(*precision*)以及它们的调和平均值 *F1-score* 作为衡量标准。

图 10 是四种方法的检测结果展示。从图中可以看出，DBSCAN-CNV 在大多数情况下的 *F1-score* 的值都是最大的，例如当数据的 *noise*(高斯噪声水平)=0.4，*purity*(肿瘤纯度)=0.3 时，信号数据的各类噪声水平是最大的，相应的检测难度也是最大的，DBSCAN-CNV 的检测结果依旧有 0.683 的灵敏度和 0.815 的准确率，而 FLLat 的灵敏度只有 0.21，准确率只有 0.56，SAIC 的灵敏度虽然有 0.79，高于 DBSCAN-CNV，

但准确率只有 0.27,所以 $F1\text{-}score$ 依旧远小于 DBSCAN-CNV,PLA 在这组数据的灵敏度和准确率为 0。

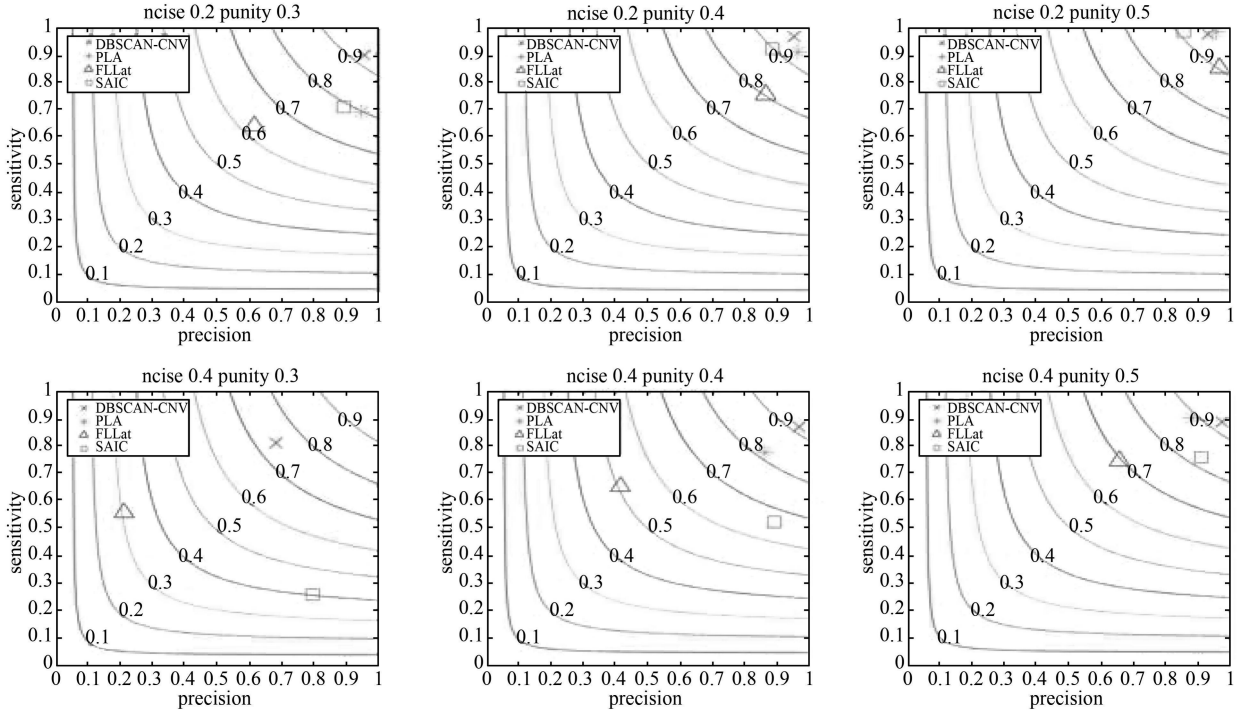

图 10 不同方法对低频率模拟数据的复发 CNV 检测结果的性能比较,标题中的 *noise* 值代表高斯噪声水平,*purity* 值代表肿瘤纯度

相较另外三种方法,DBSCAN-CNV 在噪声水平高的数据里有着明显优势,而在噪声水平较低的数据里,虽然不能保证所有结果都是最好的,但是也都有着很高的灵敏度和准确度。所以综合来看,DBSCAN-CNV 的性能稳定,表现突出,是四种方法里最优的。

2.2 真实数据

为了证明 DBSCAN-CNV 在真实数据上的可用性,本文实验将 DBSCAN-CNV 分别应用在乳腺癌真实数据和肺癌真实数据上,并将检测结果与现有研究已发现的疾病相关基因进行比对,结果证明该方法可以检测出正确的疾病相关基因。接下来是对这两种数据检测的详细介绍。

2.2.1 *乳腺癌真实数据*。该数据包含了 112 个乳腺癌样本的数据,每个样本都有 23 条染色体上的不同数据.在进行实验时,首先把不同染色体的数据分割开来,形成 23 个不同的信号数据矩阵,然后对前 22 个数据进行检测(仅在常染色体上进行检测)。由于真实数据更加杂乱无章,在检测过程中聚类的数目会比较多,因此设定阈值 $T=0.1$,如果某些类中包含点的个数加起来不超过阈值 T,则认为这些类代表的是发生复发 CNV 的位点。

以已报道的文献为标准,DBSCAN-CNV 在 9 号染色体上检测出的与乳腺癌相关的基因如表 1 所示。例如 DBC1 基因,文献[17]中说明了缺失 DBC1 对于乳腺癌的影响。又例如文献[18]中说明了 MTAP 基因对人乳腺癌细胞侵袭和迁移的影响。由此可知,DBSCAN-CNV 具有从真实乳腺癌患者数据中检测出乳腺癌相关基因的能力。

表 1 DBSCAN-CNV 在 9 号染色体上检测出的与乳腺癌相关的基因

序号	基因名称	开始位置	结束位置
1	MTAP	21792635	21855970
2	TEK	27099441	27220165
3	DBC1	120968729	121171522
4	CDK5RAP2	122190968	122382258
5	TRAF1	122704493	122728994

2.2.2　肺癌真实数据。本文实验使用的肺癌真实数据[19]中共包含 371 个肺癌患者的样本数据，每个样本都包含了 23 个染色体上的所有数据。与对乳腺癌真实数据的处理类似，首先将数据按照不同染色体分割为 23 组数据，然后对不同染色体的数据分别做检测。

以已报道的文献为标准，表 2 汇总了 DBSCAN-CNV 在 14 号染色体上检测出的与肺癌相关的基因。例如 PAX9[20]曾被多项文献报道其对肺癌的影响，而 FOXA1[21]则与抑制肺癌抗肿瘤免疫力有关。由此可知，DBSCAN-CNV 可以从真实肺癌数据中检测出肺癌相关基因。

表 2　DBSCAN-CNV 在 14 号染色体上检测出的与肺癌相关的基因

序号	基因名称	开始位置	结束位置
1	BAZ1A	34291688	34414604
2	PSMA6	34831325	34856431
3	STELLAR	35910127	35911333
4	TITF1	36055353	36058654
5	NKX2-8	36118967	36121537
6	PAX9	36200656	36215621
7	FOXA1	37128940	37134240
8	SSTR1	37746955	37752019

由 DBSCAN-CNV 对以上两种真实数据的检测结果可知，该方法可以对真实数据做出有价值的分析，这对于疾病的研究是十分重要的。

3　结论

复发 CNV 对人类复杂疾病的发生发展有着重要影响，因此研究复发 CNV 对于诊断治疗这些疾病有很大意义。本文提出了一种基于聚类的可以从多样本数据中检测出复发 CNV 的算法 DBSCAN-CNV，该算法首先将原始信号数据进行平滑处理，然后提取各位点发生单样本 CNV 的比率以及各位点的幅度均值，以这两个特征作为聚类的特征；在聚类这一步，本文采用了 DBSCAN 聚类算法，该算法虽然实现简单但适用于本文的数据；最后根据聚类结果判定哪些位点发生了复发 CNV。

在本文实验中，首先将 DBSCAN-CNV 分别应用到高频和低频两种模拟数据上，其中高频数据共有 6 种不同场景，低频数据共有 10 组不同的参数选择，同时也将 PLA、SAIC、FLLat 这三种同行算法应用在这些数据上，检测结果以灵敏度和准确率作为衡量指标。实验结果表明，DBSCAN -CNV 的性能显著优于另外三种方法。然后又将 DBSCAN-CNV 分别应用在乳腺癌和肺癌真实数据集上，检测结果中发现了现有文献报道过的疾病相关基因，这表明该算法对于真实数据也具有可用性.综上所述，DBSCAN-CNV 对于复发 CNV 的检测性能有着显著提升。

针对计算复杂度，本文所提方法 DBSCAN-CNV 的时间复杂度近似为 $O(N\log N)$，在实验过程中，与其他方法相比，本文方法的运行时间较短。

在将来研究工作中，仍然存在不足以及可以进一步扩展的工作：(1) 模拟数据假设在不同样本间发生复发 CNV 的位置完全相同，事实上它们的位置可能会有细小的差异，这可能对实验结果产生一定影响。(2) 本文仅在乳腺癌和肺癌的真实数据集上进行了实验，然而随着测序技术的发展，已经积累了海量的疾病变异数据。因此下一步可以在其他疾病的真实数据集上进行实验，以期发现更多与疾病相关的 CNV，这将是十分有意义的研究。

参　考　文　献

[1] BEROUKHIM R，MERMEL C H，PORTER D，et al. The landscape of somatic copy-number alteration across human cancers[J]. Nature，2010，463(7283)：899-905.

[2] ZHANG J，FEUK L，DUGGAN G E，et al. Development of bioinformatics resources for display and analysis of copy number and other

structural variants in the human genome[J].Cytogenetic & Genome Research,2006,115(3):205-214.

[3] STANKIEWICZ P,LUPSKI J R.tructural variation in the human genome and its role in disease[J].Annu Rev Med,2010,61:437-55.

[4] COOK,JR E H,SCHERER S W.Copy-number variations associated with neuropsychiatric conditions[J].Nature,2008,455(7215):919-23.

[5] AITMAN T J,DONG R,VYSE T J,et al.Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans[J].Nature,2006,439(7078):851-5.

[6] CONRAD D F,DALILA P,RICHARD R,et al.Origins and functional impact of copy number variation in the human genome[J].Nature,2010,464(7289):704-12.

[7] BOZIC I.,ANTAL T,OHTSUKI H,et al. Accumulation of driver and passenger mutations during tumor progression[J].Proc Natl Acad Sci U S A,2010,107(43):18545-50.

[8] YUAN X,ZHANG J,YANG,Let al.Detection of significant copy number variations from multiple samples in next-generation sequencing data[J].IEEE Trans Nanobioscience,2018,17(1):12-20.

[9] NOGHABI H S,MOHAMMADI M,TAN Y H.Robust group fused lasso for multisample copy number variation detection under uncertainty[J].Iet Systems Biology,2016,10(6):229.

[10] CAI H,CHEN P,CHEN J,et al.WaveDec:A wavelet approach to identify both shared and individual patterns of copy-number variations [J].IEEE Transactions on Bio-medical Engineering,2018,65(2):353.

[11] ZHOU X W,LIU J M,WAN X,et al.Piecewise-constant and low-rank approximation for identification of recurrent copy number variations [J].Bioinformatics,2014,30(14):1943-9.

[12] NOWAK G,HASTIE T,POLLACK JR,et al.A fused lasso latent feature model for analyzing multi-sample aCGH data[J].Biostatistics,2011,12(4):776-91.

[13] YUAN X G,YU G Q,HOU X C,et al.Genome-wide identification of significant aberrations in cancer genome[J].BMC Genomics,2012,13:342.

[14] OLSHENl A B,VENKATRAMAN E S.Circular binary segmentation for the analysis of array-based DNA copy number data[J].Biostatistics,2004,5(4):557-72.

[15] TAN P N.数据挖掘导论[M].北京:人民邮电出版社(完整版),2011.

[16] DIAZ-URIARTE O M R R.Finding recurrent copy number alteration regions: a review of methods[J].Current Bioinformatics,2010,5(1):1-17.

[17] FANG Q,BELLANTI J A,ZHENG S G.Advances on the role of the deleted in breast cancer (DBC1) in cancer and autoimmune diseases [J].J Leukoc Biol,2020.

[18] 姜雨刚,张安秦,朱彩霞,等.MTAP对人乳腺癌细胞侵袭和迁移的影响[J].国际外科学杂志,2013,40(3):160-164.

[19] WEIR B A,WOO M S,GETZ G,et al.Characterizing the cancer genome in lung adenocarcinoma[J].Nature,2007,450(7171):893-8.

[20] HSU D S,CHAITANYA R ACHARYA,BALA S B,et al.Characterizing the developmental pathways TTF-1,NKX2-8,and PAX9 in lung cancer[J].Proc Natl Acad Sci U S A,2009,106(13):5312-7.

[21] LIANG J,TIAN C,ZENG Y,et al.FOXA1+ regulatory T cells: a novel T cell subset that suppresses antitumor immunity in lung cancer [J].Biochemical and Biophysical Research Communications,2019,514(1):308-315.