基于快速搜索和查找密度峰的聚类方法*

Alex Rodriguez　　Alessandro Laio

摘要

聚类分析的目的是根据数据的相似性将其分类, 它的应用范围十分广泛, 在天文学、生物信息学、文献计量学和模式识别中都有应用。基于聚类中心通常具有密度高于其近邻点, 且密度较高的点之间的距离较大这一特征, 我们提出了一种聚类方法。在聚类过程中, 这种方法会自动确定聚类中心的数量, 同时, 离群点也会被自动剔除。无论数据分布的形状和它们所嵌入的空间维度如何, 这种算法都能够起到良好的作用。我们还在几个数据集上验证了该算法的性能。

*翻译自 Rodriguez et al.(2014) 一文[1]。

聚类算法试图根据数据之间的相似性将其分类或聚类. 目前已经有很多种不同的聚类算法[2], 但它们甚至在如何实施聚类这个问题上也有很大的不同做法. 在 K-means[3] 和 K-medoids[4] 方法中, 通常是以数据点到某个聚类中心的距离较小为一类的特征. 此时, 目标函数通常是到一组假定的聚类中心的距离之和, 然后对聚类中心进行优化[4-7], 直到找到最好的簇中心. 然而, 由于数据点总是被分配到离他最近的聚类中心, 这些聚类方法无法应用于非球形簇[8]. 而在基于概率分布的算法中, 人们试图将观察到的数据点表示为某些函数的概率分布[9], 这种方法的准确性则取决于选取的概率函数表示数据的能力.

而具有任意形状簇的数据分布很容易被基于局部密度的聚类方法检测到. 在 DBSCAN[10] 算法中, 我们需要选择一个密度阈值, 将密度低于这个阈值的区域中的点作为噪声丢弃掉, 将高密度的不相连的区域划分到不同的簇中. 但如何选择合适的阈值是一个比较困难的问题. 然而这个缺点在均值平移算法中得到了克服[11-12], 在这个算法中, 簇被定义为一组接近于密度分布函数的某一局部最大值的点. 这种方法可以找到非球形聚类, 但只适用于由一组坐标定义的数据, 且计算成本较高.

因此, 我们提出了一种替代方法. 它与 K-medoids 方法类似, 以数据点之间的距离作为聚类特征. 像 DBSCAN 和均值平移法一样, 它也能够检测到非球形聚类, 并自动确定聚类中心的数量. 与均值移动法一样, 聚类中心被定义为数据密度函数的局部最大值. 然而, 与均值移动法不同的是, 它不需要将数据嵌入到一个向量空间, 也不需要具体的将每个数据点的密度最大化.

这个算法的基础是假设聚类中心周围有局部密度较低的近邻点, 并且它们与任何局部密度较高的点都有比较大的距离。对于每个数据点 i, 我们计算两个量: 它的局部密度 ρ_i 和它与密度较高的点的距离 σ_i, 这两个量都只取决于数据点之间的距离 d_{ij}, 假设这个距离满足三角不等式, 则数据点 i 的局部密度 ρ_i 被定义为

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

其中, 如果 $x < 0$, 则 $\chi(x) = 1$, 否则 $\chi(x) = 0$, 而 d_c 是截止距离。基本上可以说是 ρ_i 等价于以 d_c 为半径的邻域内点的数量。该算法只对不同点的 ρ_i 的相对大小敏感, 这意味着, 对于大数据集, 其结果对 d_c 的选择是稳健的。

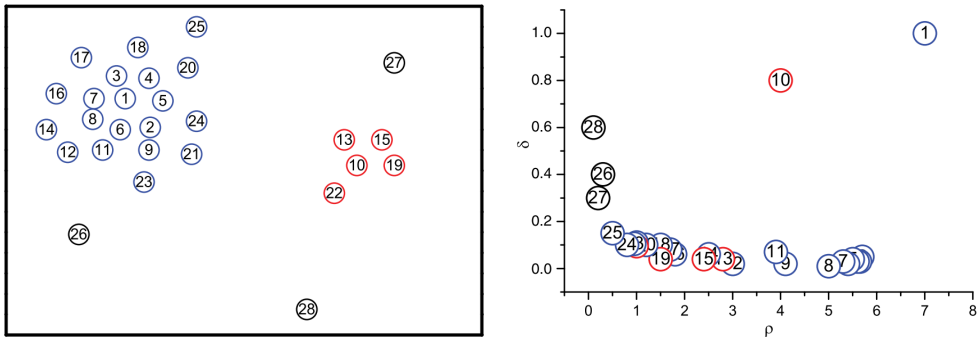
σ_i 是通过计算点 i 和任何其他密度较高的点之间的最小距离来测量的:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{2}$$

对于密度最高的点, 习惯上取 $\sigma_i = \max_j(d_{ij})$。需要注意的是只有对于密度为局部或全局最大值的点, σ_i 才会比典型的最近邻距离大得多。因此, 聚类中心被认为是 σ_i 异常的点。

捕捉这个突变, 是算法的核心步骤, 图 1 中的简单例子就可以说明, 图 1(a) 是在二维空间中的 28 个点, 我们发现密度最大的点为 1 和 10, 将其确定为聚类中心。图 1(b) 是每个点的 σ_i 与 ρ_i 的关系图, 我们将把它称为决策图。可以看到点 9 和 10 有相近的 ρ 值, 但却有差距很大的 σ 值, 而点 9 属于点 1 的簇, 其他几个 ρ 值较大的点离它很近, 而密度较高的点 10 属于另一个簇。因此, 只有具有较高 σ 和相对较高的 ρ 的点才是聚类中心。26、27、28 点的 σ 较高, 但 ρ 相对较低, 因为它们基本上是孤立的, 因此可以认为它们是由单个点组成的簇, 即离群点。

在确定聚类中心后, 其余的点都会被分配到离其最近的簇中。与其他聚类算法不同的是, 此处的聚类分配是一步完成的, 其目标函数是迭代优化的[3, 9]。



(a) 数据点的分布，按密度递减的顺序排列。 (b) ((a)) 中数据点的决策图，不同的颜色对应不同的簇。

图 1: 算法在二维数据上的示例。

在聚类分析中，定量地衡量数据点的分配的可靠性往往是十分有用的，在基于函数优化的方法中[3, 9]，其收敛时的值也是对于可靠性的一种自然度量。在像 DBSCAN[10] 这样的方法中，人们会认为密度值高于某一阈值的点是可靠的，这可能会导致低密度的聚类中心都被归为噪声，如图 2(e) 所示。在我们的算法中，我们并没有引入噪声信号的截止值，相反，我们首先为每个聚类找到一个边界区域，定义为属于该类但与属于其他类的点的距离小于 d_c 的点的全体，然后，我们找出边界区域内密度最高的点，用 ρ_b 表示其密度，簇中密度高于 ρ_b 的点被认为是簇核的一部分 (稳健分配)，其他的点被认为是簇环的一部分 (可以被认为是噪声)。

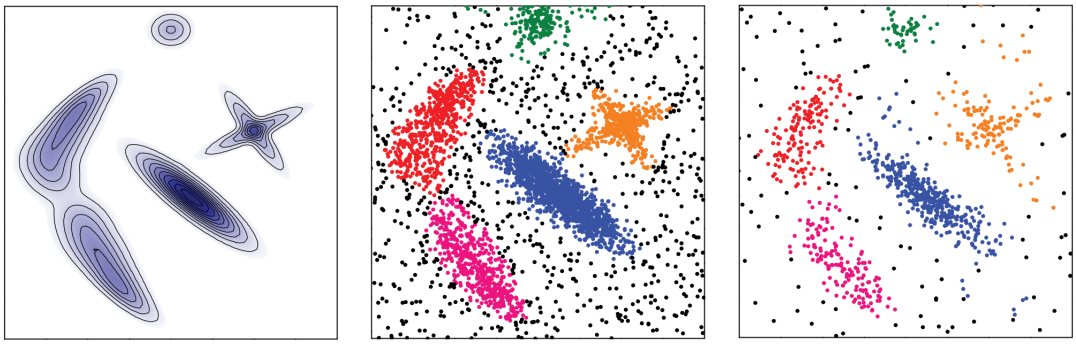
为了测试此算法，我们首先考虑图 2 中的例子，其数据是从一个具有非球形和剧烈重叠的概率分布中抽取的 (图 2(a))。图 2(b)、(c) 是，分别从图 2(a) 中中抽取的 4000 和 1000 个点，在相应的决策图 (图 2(d), (e)) 中，只看到 5 个点的 σ 值很大，密度也相当大。这些点在图中表示为大的实心圆，对应于聚类中心。选定聚类中心后，每个点要么被分配到一个簇，要么被分配到环，即使是对那些密度非常不同的 (图 2(c) 中的蓝色和浅绿色点) 和非球形峰，该算法也可以捕捉到概率分布的位置和形状。此外，通过目测图 2(a) 中的概率分布可以知道，分配到环的点对应的区域不会被分到任何类中。

为了表明该算法对大量数据也具有鲁棒性，我们从图 2(a) 中抽取 10000 个点进行分析，将从 10000 个样本上获得的聚类结果作为参考，通过只保留一部分点来获得缩小的样本，并对每个缩小的样本独立地进行聚类。图 2(f) 是分配到一个簇的点的误分率与样本量的关系图像，可以看到，即使对只包含 1000 个点的小样本，被错误分类的点占比仍然远低于 1%。

对图 2(b) 中的数据改变 d_c，会产生一致的结果 (图 S1 所示[1])。可以经验地按照如下规律选择 d_c，使平均邻域数约为数据量的 $1 \sim 2\%$。对于由少量点组成的数据集，ρ_i 可能会受到较大的统计误差的影响。在这些情况下，最好用更精确的方法估计密度[11-12]。

接下来，我们在图 3 的例子上对算法进行了测试。对于计算点少的密度，我们采用了 Cheng(1995) 中描述的指数核算法，在图 3(a) 中，我们使用了 Gionis et al.(2007) 中的一个数据集，得到的结果与

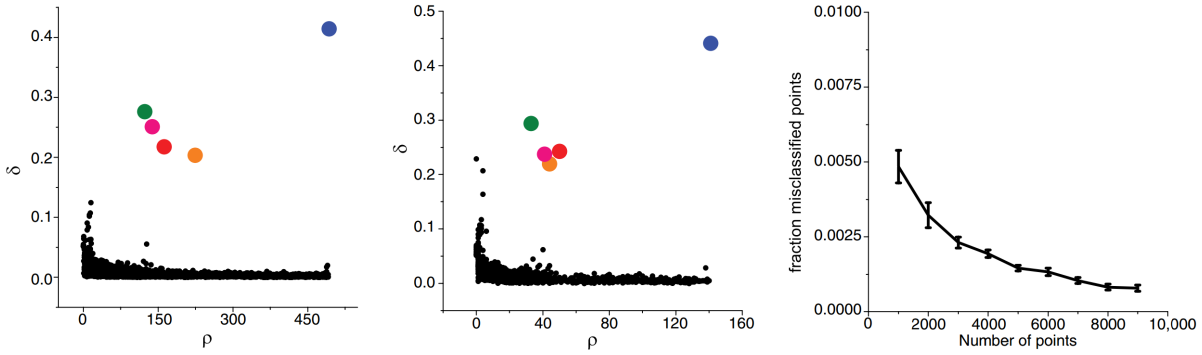
[1]类似于图 S1, S2 的此类图片可以在http://science.sciencemag.org/content/suppl/2014/06/25/344.6191.1492.DC1中获得。



(a) 数据的概率分布图示, 强度最低的区域对应 20% 的概率。

(b) 样本量为 4000 的数据分布结果, 黑色为噪声。

(c) 样本量为 1000 的数据分布结果, 黑色为噪声。



(d) 样本量为 4000 的数据的决策图。

(e) 样本量为 1000 的数据的决策图。

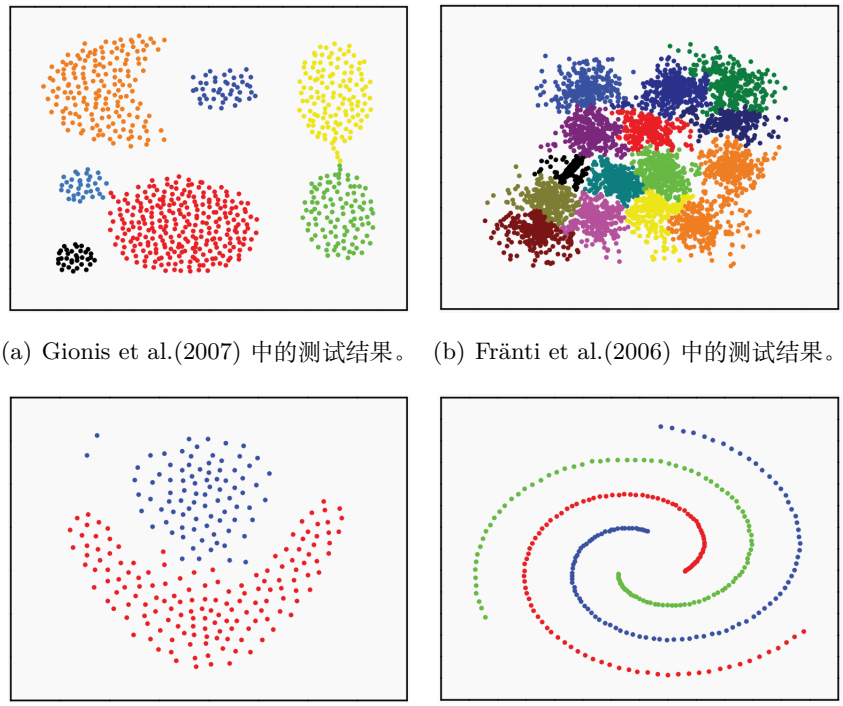
(f) 样本量和误分率的关系图, 误差条表示平均值的标准误差。

图 2: 算法在人工合成数据集上的示例。

原文的结果基本一致, 而文中说到许多常用的方法无法成功将此数据集聚类。在图 3(b) 中, 我们使用了 Fränti et al.(2006) 中的一个 15 个概率分布高度重合的例子, 我们的算法成功地确定了每个簇的结构。在图 3(c) 中, 我们使用了 FLAME 方法[15] 中的例子, 得到的结果与原始方法高度一致。在图 4(d) 所示的为了说明基于路径的谱聚类[16] 的性能而引入的数据集中, 我们的算法在不需要生成连接图的情况下, 正确地找到了三个簇。作为比较, 在图 S3 和 S4 中, 我们展示了基于路径的谱聚类[16] 的性能。图 S3 和 S4 中, 我们展示了这四个测试案例和图 2 中的例子通过 K-means[3] 获得的聚类结果。在大多数情况下, 即使使用正确的 K 值进行 K-means 聚类, 也不能得到很好的结果。

该方法对于度量的变化是稳健的, 即保持式 1 中的密度估计不变, 这些变化也不会对小于 d_c 的距离产生显著影响。显然, 式 2 中的距离会受到这种度量变化的影响, 但决策图的结构 (特别是 d 值大的数据点的数量) 是密度值排序的结果, 而不是远处的点之间实际距离的结果, 证明这一说法的例子如图 S5 所示。

我们的方法只需要测量 (或计算) 所有数据点之间的距离, 而不需要对概率分布[9] 或多维密度函数[11] 进行参数化。因此, 其性能不受数据点所嵌入的空间的内在维度影响。同时, 我们还验证了在 256 个维度[17] 的 16 个聚类的例子中, 该算法可以找到正确的聚类中心的数量, 并正确的将数据点划分到每一个簇中 (图 S6)。对于[18] 中三类小麦种子的 7 个 X 射线特征的 210 个测量数据, 该算法正



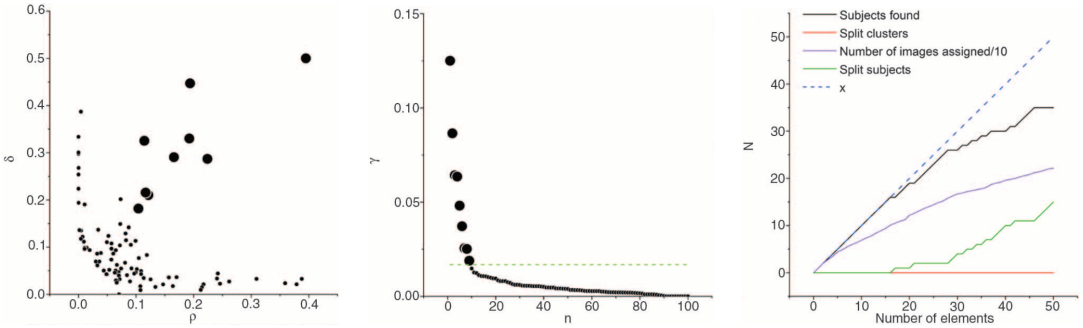
(a) Gionis et al.(2007) 中的测试结果。 (b) Fränti et al.(2006) 中的测试结果。

(c) Fu et al.(2007) 中的测试结果。 (d) Chang et al.(2008) 中的测试结果。

图 3: 其他文献中测试例子的结果。

确预测了 3 个簇的存在, 并对 97% 的数据进行了正确分类 (图 S7, S8)。

我们还将该方法应用于 Olivetti 人脸数据库[19], 这是机器学习算法的一个较为普遍的测试数据及, 使用这个数据集的目的是在没有任何先前训练的情况下, 希望算法能够识别数据中的不同的人的数量。这个数据集对我们的方法提出了严峻的挑战, 因为 “理想” 的聚类中心数量与数据集中的样本数量相同, 这使得对密度的估计变得困难。两幅图像之间的相似度是通过[20] 中的方法计算的, 密度是由一个方差为 $d_c = 0.07$ 的高斯核[12] 估计出来的。对于这样一个小的集合, 密度估计器不可避免地会产生较大的误差, 因此, 我们对于数据点的划分应该更严格一些。只有当一个图像的距离小于 d_c 时, 它才会被分配到与其最近的密度较高的图像的同一簇中。因此, 距离任何其他密度较高的图像比 dc 更远的图像会不被分配。在图 4 中, 我们显示了对数据集中前 100 张图像进行分析的结果。决策图 (图 4(a)) 显示了几个不同的密度最大值的存在。与其他例子不同的是, 它们的确切数量并不清楚, 这是因为数据比较稀疏性的结果, 按递减顺序排列的 $\gamma_i = \rho_i \sigma_i$ 图提供了一个选择聚类中心数量的提示 (图 4(b)), 这张图显示, 聚类中心的数量比较多, 从第九个数据点开始异常增长, 因此, 我们使用 9 个聚类中心来进行分析。在图 4(d) 中, 我们用不同的颜色显示了这些中心对应的聚类。7 个类对应不同的主体, 表明算法能够 “识别”10 个人物中的 7 个, 第 8 个被划分到了两个不同的类中。当对数据中的所有 400 张图像进行分析时, 决策图不能清楚地识别出聚类的数量 (图 S9)。然而, 在图 4(c) 中, 可以看出通过增加聚类中心, 大约有 30 个人物可以被毫不含糊地识别出来 (图 S9), 当加入更多的中心时, 一些人物的图像会被划分到两个簇内, 但所有的簇仍然是一致的, 即只包括同一人物的图像。按照 Dueck et al.(2007) 的方法, 我们还计算了同一人物的图像被正确划分到同一簇的



(a) 数据库中前一百张图片的决策图[19]。 (b) $\gamma_i = \rho_i \sigma_i$, 其值依次递减。 (c) 性能曲线。

(d) 前 100 张图像的聚类结果的图示。具有相同颜色的面孔属于同一簇, 聚类中心用白色圆圈标注。

图 4: Olivetti 人脸数据库上的聚类分析。(c) 中, 黑线是被识别为人像的数量, 红线是包含超过一个人像的簇数, 绿线是在簇中分裂的人物数量, 紫线是划分给一个簇的图像数量除以 10。

比率 (r_{true}) 和不同人物的图像被错误地分配到同一簇的比率 (r_{false})。如果在划分中不应用 d_c 处的截止值 (即应用我们算法的一般公式), 则在约 42 到约 50 个中心的情况下, 可以得到 $r_{\text{true}} \sim 68\%$ 和 $r_{\text{false}} \sim 1.2\%$, 这一性能与无监督图像分类的最先进方法相当[21]。

最后, 我们对聚类算法进行了基准分析, 在 300K 的水[22] 中对三丙氨酸的分子动力学轨迹进行了分析, 这种情况下, 聚类将近似对应于动力学盆地, 即长时间上稳定的和且被自由能量壁垒隔绝的系统独立构象, 只有在微观时间尺度上偶尔跨越。我们首先通过标准方法[23], 基于动力学矩阵的谱分析对轨迹进行了分析, 其矩阵特征值与系统的松弛时间有关。在第七个特征值 (图 S10) 之后存在一个空隙, 表明该系统有八个盆地, 与此相一致, 我们的聚类分析 (图 S10) 也产生了八个聚类, 这与动力学盆地的构象一一对应[23]。

识别具有密度最大值的聚类是一个简单而直观的选择, 就像这里和其它基于密度的聚类算法[10-11] 所做的那样, 但这种方法有一个重要的缺点, 如果随机生成数据点, 那么对于有限样本量所估计的密度远远不是均匀的, 而是以几个最大值为特征。然而, 决策图允许我们将真正的聚类与噪声产生的密度波纹区分开来。定量来看, 只有在前一种情况下, 对应于聚类中心的点与其它点在 ρ 和 σ 上有相当大的差距, 对于随机分布, 人们反而会观察到 ρ 和 σ 的连续分布。事实上, 我们对超立方体中的均匀分布随机生成的点进行了分析, 在式 1 和 2 的点之间的距离是在超立方体上的周期性边界条件

下计算出来的, 这一分析表明, 对于随机分布的数据点, 数量 $\gamma_i = \rho_i\sigma_i$ 是按照幂律分布的, 其指数取决于点所在空间的维度。对于实际的数据集, 如图 2 至图 4 中的数据集, γ 的分布与幂律有明显的不同, 特别是在高 γ 的区域 (图 S11) 上。这一结果可以为自动选择聚类中心的提供标准以及在统计学上对我们的算法的验证的可靠性提供依据。

参考文献

[1] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[2] XU R, WUNSCH D. Survey of clustering algorithms[J]. IEEE Transactions on neural networks, 2005, 16(3): 645-678.

[3] MACQUEEN J, et al. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability: vol. 1: 14. 1967: 281-297.

[4] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis[M]. John Wiley & Sons, 2009.

[5] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[6] WARD JR J H. Hierarchical grouping to optimize an objective function[J]. Journal of the American statistical association, 1963, 58(301): 236-244.

[7] HÖPPNER F, KLAWONN F, KRUSE R, et al. Fuzzy cluster analysis: methods for classification, data analysis and image recognition[M]. John Wiley & Sons, 1999.

[8] JAIN A K. Data clustering: 50 years beyond K-means[J]. Pattern recognition letters, 2010, 31(8): 651-666.

[9] MCLACHLAN G J, KRISHNAN T. The EM algorithm and extensions[M]. John Wiley & Sons, 2007.

[10] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise.[C]//Kdd: vol. 96: 34. 1996: 226-231.

[11] FUKUNAGA K, HOSTETLER L. The estimation of the gradient of a density function, with applications in pattern recognition[J]. IEEE Transactions on information theory, 1975, 21(1): 32-40.

[12] CHENG Y. Mean shift, mode seeking, and clustering[J]. IEEE transactions on pattern analysis and machine intelligence, 1995, 17(8): 790-799.

[13] GIONIS A, MANNILA H, TSAPARAS P. Clustering aggregation[J]. Acm transactions on knowledge discovery from data (tkdd), 2007, 1(1): 4-es.

[14] FRÄNTI P, VIRMAJOKI O. Iterative shrinking method for clustering problems[J]. Pattern Recognition, 2006, 39(5): 761-775.

[15] FU L, MEDICO E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data[J]. BMC bioinformatics, 2007, 8(1): 1-15.

[16] CHANG H, YEUNG D Y. Robust path-based spectral clustering[J]. Pattern Recognition, 2008, 41(1): 191-203.

[17] FRANTI P, VIRMAJOKI O, HAUTAMAKI V. Fast agglomerative clustering using a k-nearest neighbor graph[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(11): 1875-1881.

[18] CHARYTANOWICZ M, NIEWCZAS J, KULCZYCKI P, et al. Complete gradient clustering algorithm for features analysis of x-ray images[G]//Information technologies in biomedicine. Springer, 2010: 15-24.

[19] SAMARIA F S, HARTER A C. Parameterisation of a stochastic model for human face identification[C]//Proceedings of 1994 IEEE workshop on applications of computer vision. 1994: 138-142.

[20] SAMPAT M P, WANG Z, GUPTA S, et al. Complex wavelet structural similarity: A new image similarity index[J]. IEEE transactions on image processing, 2009, 18(11): 2385-2401.

[21] DUECK D, FREY B J. Non-metric affinity propagation for unsupervised image categorization[C]//2007 IEEE 11th International Conference on Computer Vision. 2007: 1-8.

[22] MARINELLI F, PIETRUCCI F, LAIO A, et al. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations[J]. PLoS Comput Biol, 2009, 5(8): e1000452.

[23] HORENKO I, DITTMER E, FISCHER A, et al. Automated model reduction for complex systems exhibiting metastability[J]. Multiscale Modeling & Simulation, 2006, 5(3): 802-827.