环,否则,r 加 1 继续进行计算.在循环结束后,由于我们得到的 $NN_r(i)$ 个数可能太大,以至于不好分开两个靠近的类别簇,这里,我们取前 2/3 个邻域点作为新的邻域进行扩张.在对每个点分类之前,首先判定 $NN_r(i)$ 邻域内的点是否已被分类,如果存在已经被分类的点,那么选取邻域内最近的被分类的点的标签赋给点 i;如果不存在,则把当前标签 Lay 赋值给点 i.对点 i 分类完成后,将点 $Nb(i)$ 赋值为–1,因为每次要寻找 $Nb(i)$ 最大的点,所以保证一些 $Nb(i)=0$ 的离散点也能被找到.之后把点 i 的 $NN_r(i)$ 邻域放在一个队列,依次计算队列内未被分类的点,如果第 1 个点未被分类,则按照上面给点 i 分类的方法进行分类,并将第 1 个点的未被分类的邻域内的点放在队列中,移除第 1 个点.直到队列内的点全部分类完毕,即队列清空,表示一个类簇形成,类别数 Lay 加 1.

算法 2. CDD 算法.

输入:数据集 X;

输出:C:类别标签.

主要步骤:

1. 初始化: $r=1, Nb(i)=0, NN_r(i)=\varnothing$;
2. 分别对 X 中的每个点 i 计算它的第 r 近邻 j;
   a. $Nb(j)=Nb(j)+1$
   b. $NN_r(i)=NN_r(i)\cup\{j\}$
3. 计算 $Nb(i)=0$ 的点的数量 Num;
4. 如果 Num 的值不变:
   $r=r+1$, 重新计算第 2 步和第 3 步;
   如果 Num 的值仍然不变:
   进入步骤 5;
   否则:
   进入步骤 2;
   否则:
   $r=r+1$, 进入步骤 2;
5. 保存并输出 $NN_r(i), Nb(i)$;
6. 取前 $2\times|NN_r(i)|/3$ 个点作为新的 $NN_r(i)$;
7. **While**(**1**):
8. 找到 $Nb(i)$ 最大的点 x;
9. 如果 $y\in NN_r(x)$ 且 $Class(y)\neq 0$:
10. 那么把距 x 最近的 y 的类标签赋给 x;
11. 否则:
12. $Class(x)=Lay$;
13. 把 $NN_r(i)$ 内的点放进一个空队列;
14. **While**(队列非空):
15. 选择队列第 1 个点 x_p;
16. 如果 $Nb(x_p)\neq -1$ (未分类):
17. 如果 $y'\in NN_r(x_p)$ 且 $Class(y')\neq 0$:
18. 那么把距 x_p 最近的 y'的类标签赋给 x_p;
19. 否则:
20. $Class(x_p)=Lay$;
21. $Nb(x_p)=-1$;
22. 把 $NN_r(x_p)$ 内未分类的点插入到队列中;

23. 移除队列内的第 1 个点;
24. **end While**
25. $Lay = Lay + 1$;
26. 如果所有点 $Nb(i) = -1$ 成立:
27. 那么 **Break**;
28. **end While**

3 实验结果及分析

为了验证本文所提 CDD 算法的有效性,我们进行了仿真实验.实验中,采用 4 组带噪声的人工数据集,对算法的性能进行比较.实验中,我们将 IS-DBSCAN、DBSCAN 和本文所提算法 CDD 进行了对比.实验结果表明,本文所提 CDD 算法可以有效地确定数据集的类别数,实现对含有噪声数据集的有效分类.实验中所采用的数据集如图 8 所示,分别为 *D*2、*D*3、*D*4 和 *D*5.实验对比结果如图 9 所示.

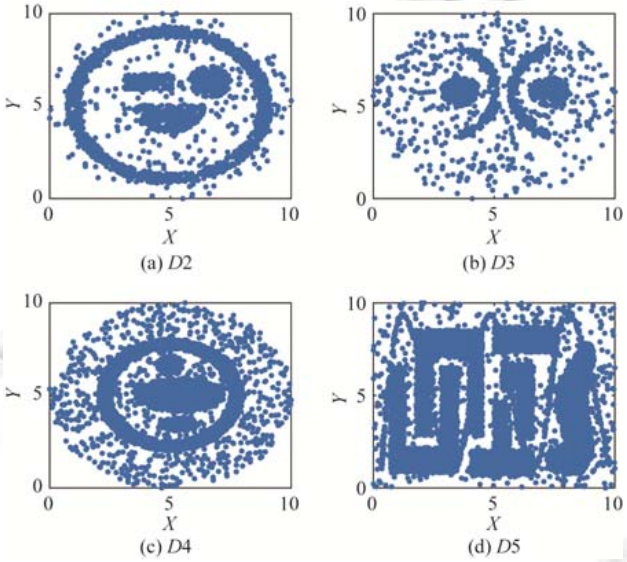


Fig.8 Four groups of datasets used in the experiments

图 8 实验所使用的 4 组数据集

由图 9 所示结果可以看出:对于数据集 *D*2,3 种算法都能检测出离散点正确地分为 5 类.对于数据集 *D*3,由于 IS-DBSCAN 算法去除噪声的效果不好,很多噪声没有去除,最终把数据分为 9 类,原因是把其中一些离散点当作了有用点;DBSCAN 算法和 CDD 算法能够检测离散点,正确地分为了 5 类.对于数据集 *D*4,IS-DBSCAN 分成了 9 类,除了正确地分出 5 类数据点外,还包括 4 个错把噪声当成有用点的类簇;DBSCAN 算法和 CDD 算法正确地分成了 5 类.对于数据集 *D*5,IS-DBSCAN 算法最终分成了 11 类,DBSCAN 算法结果分成了 8 类,只有 CDD 算法正确地分为了 7 类,它不会把靠近有用点的噪声点单独分为一类.综上分析,对比使用两个参数的 DBSCAN 算法和使用一个参数的 IS-DBSCAN 算法,CDD 算法在只使用一个参数的情况下,能够实现噪声的检测和得出正确的类别数,而不会出现把一些未被检测到的离散点单独归为一类的情况.

为了进一步验证所提算法的有效性,我们对算法所得聚类质量作进一步对比.这里,我们采用平均聚类纯度[19]对算法性能进行定量分析,其定义如下:

$$pur = \frac{\sum_{i=1}^{K} \frac{|C_i^d|}{|C_i|}}{K} \times 100\% \tag{3}$$

其中, K 表示聚类的数量,$|C_i^d|$ 表示在类别 i 中占主导地位类别标签的点的数量.$|C_i|$ 表示类别 i 中点的数量.这里对于少数离散点单独分为一类的情况,将 $|C_i^d|$ 视为 0.直观来看,纯度值越大,则表示类别个数分类准确且聚类质量越高;纯度值越小,则表示类别个数越多且聚类质量越差.

3 种算法对 4 组数据集纯度计算结果见表 1.由表 1 所示结果可以看出,对于数据集 *D*2,3 种算法的 *pur* 相同;对于数据集 *D*3、*D*4 和 *D*5,可以看出本文算法的优势,CDD 算法的 *pur* 值高于其他算法.相对于 DBSCAN 算法使用两个参数,CDD 算法仅使用一个参数.

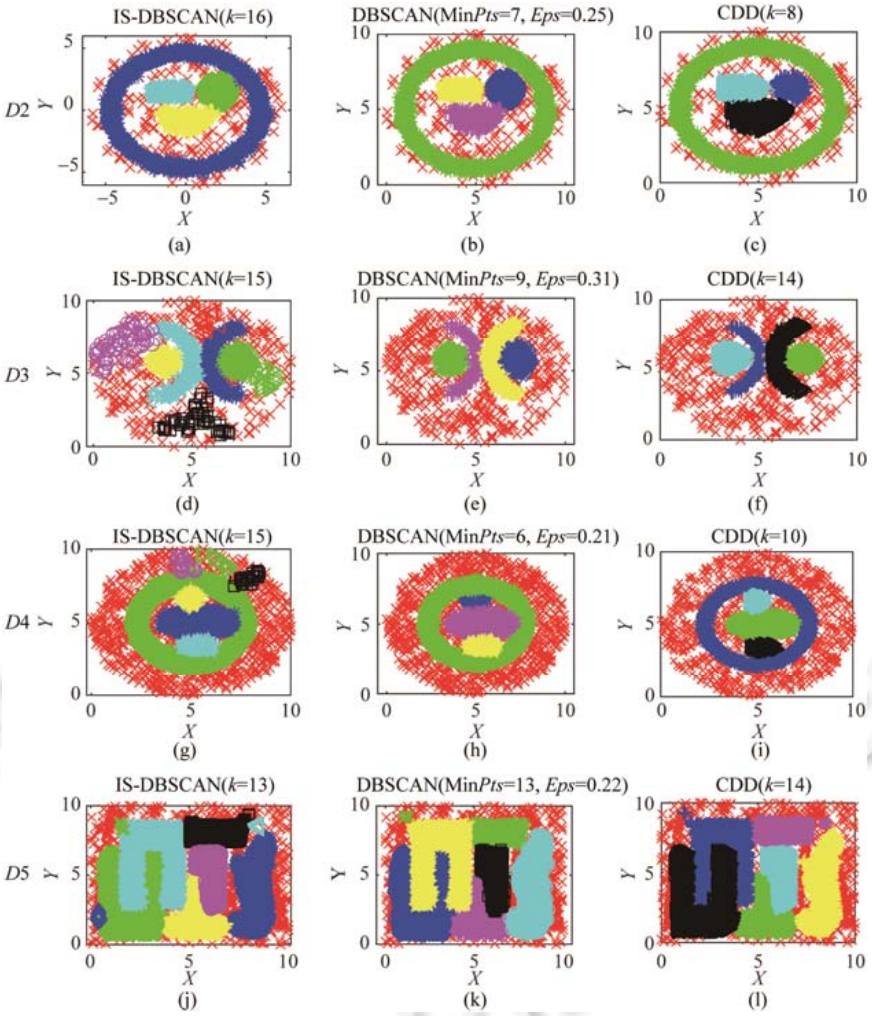


Fig.9 Clustering comparison results by IS-DBSCAN, DBSCAN and CDD for four datasets. Clusters arerepresented by different colors

图 9 IS-DBSCAN、DBSCAN 和 CDD 3 种算法在 4 组数据集上的对比结果,不同颜色代表不同类别

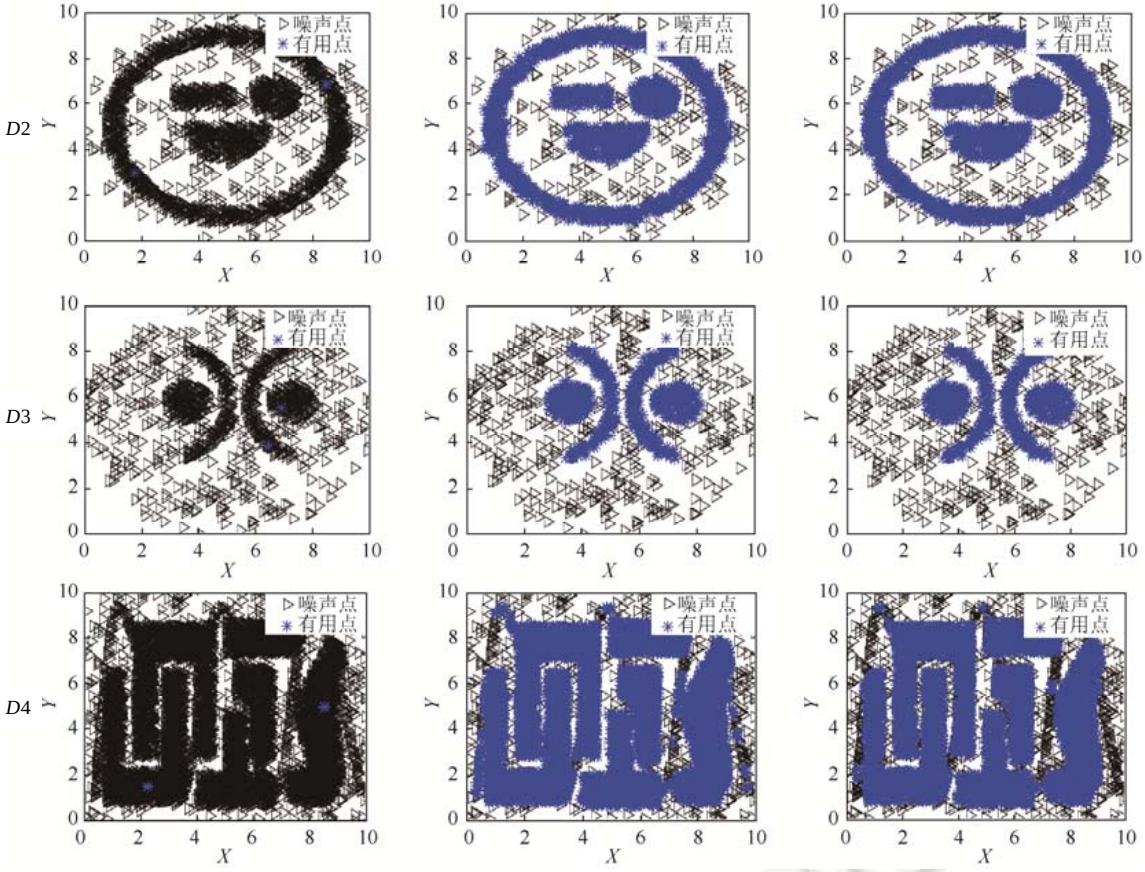
Table 1 Purity calculation results

表 1 纯度计算结果

数据集	IS-DBSCAN	DBSCAN	CDD
*D*2	96%(5 类)	96%(5 类)	96%(5 类)
*D*3	49%(9 类)	87%(5 类)	91%(5 类)
*D*4	51%(9 类)	92%(5 类)	96%(5 类)
*D*5	64%(11 类)	85%(8 类)	99%(7 类)

为了说明算法中一些参数的经验取值问题和 k 值敏感性问题,我们进行了实验对比分析.

(1) 针对噪声检测部分的取值 5%~95%,分别取全部点、5%~95%及 10%~90%这 3 种情况对 $D2$、$D3$、$D5$ 进行噪声检测,实验对比结果如图 10 所示,三角形区域所示为检测的噪声点,*符号区域所示为有用数据,这里,k 默认为 10.从对比结果可以看出,如果我们不设定范围,从全部点中寻找阈值,那么效果非常糟糕,几乎所有的点都判定成了噪声;从设定 5%~95%、10%~90%两个范围来看,可以很好地分离出噪声和有用点,对噪声的检测具有鲁棒性.



(a) (b) (c)

Fig.10 Noise detection results of different data range.

(a) column: All points, (b) column: 5%~95%, (c) column: 10%~90%

图 10 不同数据点范围检测噪声结果,(a)列:全部点,(b)列:5%~95%,(c)列:10%~90%

(2) 针对自动分类过程中,取前 2/3 个邻域点作为每个点新邻域问题,我们取 $D2$、$D3$、$D4$、$D5$ 进行实验对比分析,分别设置邻域系数为 1/2、2/3、5/6 和 1.不同邻域系数下得到的最终聚类数目见表 2,从表中可以看出,在不同邻域系数下,聚类数目变化不够明显.聚类结果的 *pur* 值如图 11 所示.从结果可以看出:当邻域系数为 1/2 时,$D3$ 和 $D4$ 数据集分类效果不好;当邻域系数为 1 时,相对比较密集的 $D5$ 数据集分类效果不好;邻域系数为 2/3 和 5/6 时都能够得出正确的类别数.这里,我们选取 2/3 作为系数是因为对于一些比较密集的数据集效果会更好.

(3) 对参数 k 敏感性的分析,我们对数据集 $D2$、$D3$、$D4$ 和 $D5$ 分别从 k=6 到 k=20 之间间隔为 2 取值进行对比实验.表 3 是聚类数目随着 k 值变化的结果,可以看出,$D2$、$D3$ 和 $D4$ 数据集在一定区间内是稳定的,$D5$ 数据集则波动较大.*pur* 值的计算结果如图 12 所示.从结果可以看出,虽然 k 变化很大,但对于数据集 $D2$ 没有影响,数据集 $D3$ 和 $D4$ 相对波动较小,而相对比较密集的 $D5$ 数据集波动较大,但只要设置正确的 k 值,也能分类出

准确的结果.

Table 2 Comparison of clustering numbers of different neighborhood coefficients

表 2 不同邻域系数聚类数目对比

数据集	1/2	2/3	5/6	1
*D*2	5	5	5	5
*D*3	5	5	5	5
*D*4	4	5	5	5
*D*5	7	7	7	6

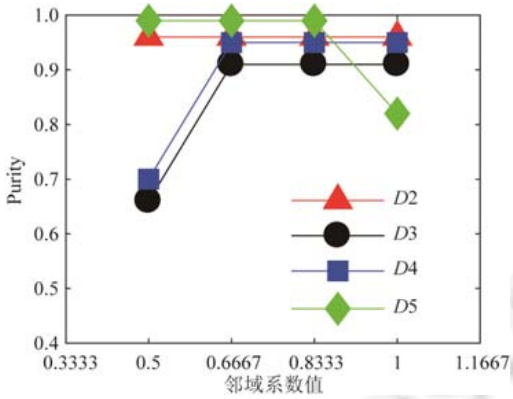


Fig.11 The clustering results obtained with different neighborhood coefficients

图 11 不同邻域系数聚类变化情况

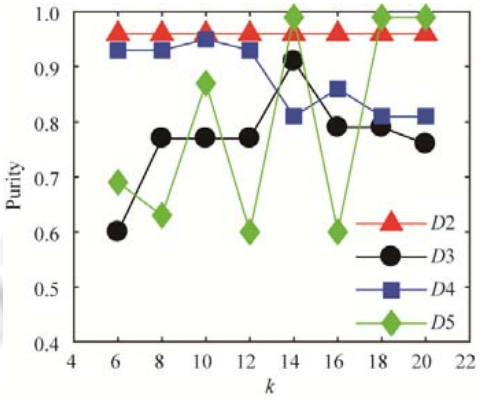


Fig.12 Clustering results obtained with different *k* values

图 12 不同 *k* 值聚类变化情况

Table 3 Comparison of clustering numbers of different *k* values

表 3 不同 *k* 值聚类数目对比

数据集	6	8	10	12	14	16	18	20
*D*2	5	5	5	5	5	5	5	5
*D*3	8	6	6	6	5	4	4	6
*D*4	5	5	5	5	2	3	2	2
*D*5	11	7	8	5	7	5	7	7

4 总结与展望

本文提出了一种基于密度差分的自动聚类算法,算法根据噪声数据与有效数据密度分布的差异,提出了基于密度差分的噪声点检测,并通过构建邻域利用近邻的方法,进一步实现了对有用数据类别的自动划分.所提 CDD 算法在实现聚类的过程中仅需要输入近邻参数 *k*,无需提前设定类别数,即可实现对数据集类别数和类别的自动划分,因此,在实际中具有更广泛的应用前景.实验验证了所提算法的有效性.未来工作将致力构造更合理的密度函数,实现自动确定近邻参数.

References:

[1] Kaufman L, Rousseeuw PJ. Finding groups in data: An introduction to cluster analysis. DBLP, 1990. http://dblp.org/rec/books/wi/KaufmanR90.html [doi: 10.1002/9780470316801]

[2] Han J, Kamber M. Data Mining: Concepts and Techniques. Burlington: Morgan Kaufmann Publishers, 2001.

[3] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys, 1999,31:264–323. [doi: 10.1145/331499.331504]

[4] Kriegel HP, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. on Knowledge Discovery from Data, 2009,3(1):1–58. [doi: 10.1145/1497577.1497578]

[5] Sun JG, Liu J, Zhao LY. Clustering algorithms research. Ruan Jian Xue Bao/Journal of Software, 2008,19(1):48–61 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.3724/SP.J.1001.2008.00048]

[6] Ng RT, Han J. CLARANS: A method for clustering objects for spatial data mining. IEEE Trans. on Knowledge & Data Engineering, 2002,14(5):1003–1016. [doi: 10.1109/TKDE.2002.1033770]

[7] Liang J, Bai L, Cao F. *K*-Modes clustering algorithm based on a new distance measure. Journal of Computer Research and Development, 2010,47(10):1749–1755 (in Chinese with English abstract). http://crad.ict.ac.cn/EN/Y2010/V47/I10/1749.html

[8] Wang W, Yang J, Muntz R. Sting: A statistical information grid approach to spatial data mining. In: Jarke M, Carey MJ, Dittrich KR, Lochovsky FH, Loucopoulos P, Jeusfeld MA, eds. Proc. of the 23rd Int'l Conf. on Very Large Data Bases. San Francisco: Morgan Kaufman Publishers, 1997. 186–195. http://www.vldb.org/conf/1997/P186.PDF

[9] Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering points to identify the clustering structure. ACM SIGMOD Record, 1999,28(2):49–60. [doi: 10.1145/304182.304187]

[10] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han JW, Fayyad U, eds. Proc. of the KDD, Vol. 96. AAAI Press, 1996. 226–231. https://aaai.org/Library/KDD/1996/kdd96-037.php

[11] Hinneburg A, Keim DA. An efficient approach to clustering in large multimedia databases with noise. In: Agrawal R, Stolorz P, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998. 58–65. http://www.ub.uni-konstanz.de/kops/volltexte/2008/7049/

[12] Ester M, Kriegel HP, Xu X. Density-Based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data Mining & Knowledge Discovery, 1998,2(2):169–194. [doi: 10.1023/A:1009745219419]

[13] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability, Vol. 1. Berkeley: University of California Press, 1967. 281–297. https://projecteuclid.org/euclid.bsmsp/1200512992

[14] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. ACM, 1996. 103–114. [doi: 10.1145/235968.233324]

[15] Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases. ACM SIGMOD Record, 1998,26(1):35–58. [doi: 10.1145/276305.276312]

[16] Karypis G, Han EH, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. Computer, 2002,32(8):68–75. [doi: 10.1109/2.781637]

[17] Cassisi C, Ferro A, Giugno R, Pigola G, Pulvirenti A. Enhancing density-based clustering: Parameter reduction and outlier detection. Information Systems, 2013,38(3):317–330. [doi: 10.1016/j.is.2012.09.001]

[18] Yang L, Zhu Q, Huang J, Cheng D. Adaptive edited natural neighbor algorithm. Neurocomputing, 2017,230:427–433. [doi: 10.1016/j.neucom.2016.12.040]

[19] Cao F, Ester M, Qian W, Zhou A. Density-Based clustering over an evolving data stream with noise. In: Proc. of the SIAM Int'l Conf. on Data Mining. Bethesda, 2006. 328–339. [doi: 10.1137/1.9781611972764.29]

附中文参考文献:

[5] 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1):48–61. http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.3724/SP.J.1001.2008.00048].

[7] 梁吉业,白亮,曹付元.基于新的距离度量的 *K*-Modes 聚类算法.计算机研究与发展,2010,47(10):1749–1755.

陈朝威(1994－),男,山东菏泽人,硕士生,主要研究领域为聚类算法,深度学习.

常冬霞(1977－),女,博士,副教授,CCF 专业会员,主要研究领域为模式识别,图像处理.