ゆるやかな密度変化に対応するLDBSCANの拡張

井上 健治† 上原子 正利† 小柳 滋†
†立命館大学 情報理工学部

1 はじめに

クラスタリング手法の1つであるDBSCAN[1]は，密度に基づくクラスタリングの代表的な手法である．このアルゴリズムには密度の違うクラスタを同時に生成できないという欠点がある．この点を改良したLDBSCAN[2]は，密度の違うクラスタを同時に生成できるが，密度変化がゆるやかな場合，別クラスタに分ける方が自然に見えるデータ群を1つにまとめてしまう欠点がある．

本稿ではLDBSCANを拡張し，この欠点を改善する手法を提案する．この手法は，クラスタを拡大する過程で密度の変化を監視し，密度の谷を検知すればクラスタの拡大を停止する．

2 既存手法

DBSCANは入力値としてεと$MinPts$を受け取り，半径ε以内に$MinPts$個以上の点を含む点から次々と点を探索し，その経路上の点の集合をクラスタとする．しかし，これらの値はグローバルであるため，DBSCANは密度の異なるクラスタを同時に生成することができない．

LDBSCANはDBSCANを改良し，各点についての外れ値の程度を表すLOF[3]と，その点の周辺にどれだけ点が密集しているかを表すLRDという値を導入することで，密度の異なるクラスタを同時に生成できる．以下では，点pのLRD，LOFをそれぞれ$D(p)$，$F(p)$と記す．しかし，LDBSCANは，例えば正規分布に従うデータ群が複数隣り合うデータ群のような，探索の過程でLRDがゆるやかに変化するデータ群をすべて同じ1つのクラスタとしてしまう．

3 提案手法

我々の提案手法は，ゆるやかなLRDの変化に対応するLDBSCANの拡張である．この手法はLRDの値が最も大きい点から幅優先探索を始める．この手法では2種類の探索経路，「確定経路」と「仮確定経路」を設定する．探索の過程でLRDの変化を監視し，LRDが増加しない限り確定経路として探索を進め，少しでもLRDが増加すれば，その点から仮確定経路として探索を進め，LRDが大きく増加すれば（LRDの谷を検知すれば）探索を停止する．そして、未探索の点でLRDの値が最も大きい点から同様に探索をはじめる．

幅優先探索で使うキュー (以下 OPEN) に格納するデータは1つの経路を表現する．これは，経路の終端点，経路上の最小LRD，最大LRD，経路の種類の4つからなる．以下ではこれを [終端点，D_{min}，D_{max}，経路の種類] の形式で記す．

経路tにおいて閾値$VR1$，$VR2$に対して以下を同時に満たすとき，点cへの到達によりLRDの谷ができると定義する．以下では，経路tの最小LRD，最大LRDをそれぞれ$D_{min}(t)$，$D_{max}(t)$と記す．

$$(D(c) - D_{min}(t))/(D_{max}(t) - D_{min}(t)) > VR1 \quad (1)$$
$$D_{min}(t)/D_{max}(t) < VR2 \quad (2)$$

(1)は経路上のLRDの変化の幅に対するcのLRDの比率を条件としている．また，一様分布に従うようなデータを探索する場合，LRDの変化の幅は常に小さく，(1)を満たしてしまう恐れがあるため，(2)を設ける．

経路tにおいて閾値$VR2$に対して以下を同時に満たすとき，経路tの終端点pから点cへの到着によりLRDが増加すると定義する．(4)と(2)は同じものである．

$$D(p) < D(c) \quad (3)$$
$$D_{min}(t)/D_{max}(t) < VR2 \quad (4)$$

経路tにおいて閾値pctに対して以下を満たすとき，経路tの終端点pから点cへの到着によりLRDが極端

An Extension of LDBSCAN for Gradual Density Changes
†Kenji INOUE
†Masatoshi KAMIHARAKO
†Shigeru OYANAGI
†College of Information Science and Engineering, Ritsumeikan University

に変化すると定義する [2].

$$D(p)/(1+pct) < D(c) < D(p)*(1+pct) \quad (5)$$

各点は「確定クラスタ ID」と「仮確定クラスタ ID」を変数として持つ．確定経路上の点は確定クラスタ ID に，仮確定経路上の点は仮確定クラスタ ID に記録する．初期値は共に -1 とする．値が共に -1 の点のクラスタ ID は未確定である．LOF が閾値以下の点の確定クラスタ ID の値は 0 とする．値が 0 の点はノイズを意味する．点の最終的なクラスタ ID は，確定クラスタ ID が -1 ならば仮確定クラスタ ID の値，そうでなければ確定クラスタ ID の値とする．

アルゴリズムを以下に示す．

■main(すべての点,$LOFUB$,pct,$MinPts$,$VR1$,$VR2$)

1. 現クラスタ ID を 0 とする．
2. 空の集合 $POINTS$ を用意し，すべての点を追加する．
3. $POINTS$ の各点の $MinPts$ 個の近傍点を調べる．
4. $POINTS$ の各点の LRD，LOF を計算する．
5. $POINTS$ のすべての点のクラスタ ID が未確定でなくなるまで以下を繰り返す．
   (a) $POINTS$ の点の中でクラスタ ID が未確定で LRD が最大の点を p とする．
   (b) $F(p)$ が $LOFUB$ 以下ならば以下を行う．
      i. 現クラスタ ID を 1 増やす．
      ii. expand_cluster(p, pct, 現クラスタ ID, $POINTS$, $VR1$, $VR2$)
   (c) そうでなければ以下を行う．
      i. p の確定クラスタ ID を 0 とする．

■expand_cluster(p, pct, 現クラスタ ID, $POINTS$, $VR1$, $VR2$)

1. 空のキュー $OPEN$ を用意する．
2. p の確定クラスタ ID を現クラスタ ID とする．
3. $OPEN$ の末尾に [p, $D(p)$, $D(p)$, 確定経路] を追加する．
4. $OPEN$ が空になるまで以下を繰り返す．
   (a) $OPEN$ の先頭を取り出し，t とする．
   (b) t の終端点の $MinPts$ 個の近傍点 c 全てについて以下を繰り返す．
      i. $minlrd = min(D_{min}(t), D(c))$.
      ii. $maxlrd = max(D_{max}(t), D(c))$.
      iii. c の確定クラスタ ID が 0 または -1 で，かつ LRD が極端に変化せず，かつ LRD の谷ができないならば以下を行う．
         A. t の経路の種類が確定経路かつ LRD が増加しないならば，c の確定クラスタ ID を現クラスタ ID とし，$OPEN$ の末尾に [c, $minlrd$, $maxlrd$, 確定経路] を追加する．
         B. そうでなく，c の仮確定クラスタ ID が現クラスタ ID でもなければ以下を行う．
            - c の仮確定クラスタ ID を現クラスタ ID とする．
            - $OPEN$ の末尾に [c, $minlrd$, $maxlrd$, 仮確定経路] を追加する．

4 実行結果

図 1 に LDBSCAN の実行結果，図 2 に本手法の実行結果を示す．黒の点はノイズを表す．2 つの正規分布に従うデータ群は，図 1 では 1 つのクラスタ (赤) となり，図 2 では別のクラスタ (赤と紫) となった．図 2 の結果のほうが自然なクラスタと言えるだろう．

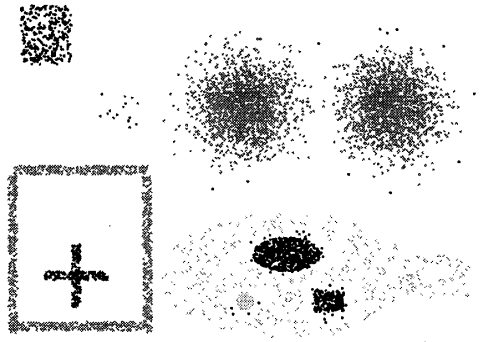

図 1 LDBSCAN の実行結果 ($MinPts = 15$, $pct = 0.25$, $LOFUB = 1.5$)

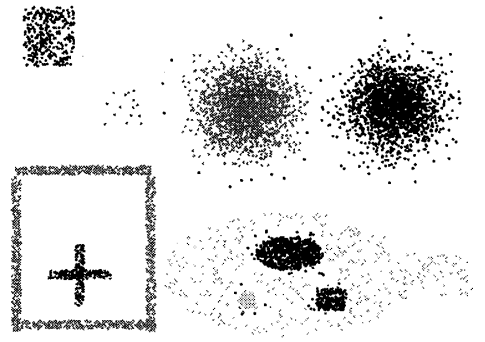

図 2 本手法の実行結果 ($MinPts = 15$, $pct = 0.25$, $LOFUB = 1.5$, $VR1 = 0.7$, $VR2 = 0.3$)

5 おわりに

本稿では LDBSCAN を拡張し，LDBSCAN の欠点であるゆるやかな密度変化に対応する手法を提案した．本手法は入力する閾値の数が多いため，今後，適切な閾値の決定方法を検討する必要がある．

参考文献

[1] Ester, M., Kriegel, H., S, J. and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, AAAI Press, pp. 226–231 (1996).

[2] Duan, L., Xu, L., Guo, F., Lee, J. and Yan, B.: A local-density based spatial clustering algorithm with noise, *Inf. Syst.*, Vol. 32, No. 7, pp. 978–986 (2007).

[3] Breunig, M. M., Kriegel, H., Ng, R. T. and Sander, J.: LOF: identifying density-based local outliers, *SIGMOD Rec.*, Vol. 29, No. 2, pp. 93–104 (2000).