DBSCAN のパラメータ設定の自動化について

慶應義塾大学大学院　理工学研究科　白井 智鵬
慶應義塾大学　理工学部　鈴木 秀男

1. はじめに

クラスタリングは，顧客分類，テキスト分類，画像解析など幅広いドメインで用いられており，その解析の目的やデータの特徴に応じて異なる手法が用いられている．このように様々な場面で利用されており，多くの種類があるクラスタリング手法だが，大きく分けると，分割型クラスタリング，階層型クラスタリング，モデルベースクラスタリング，密度ベースクラスタリング，グリッドベースクラスタリングに分類することができる．本研究のベースとなる DBSCAN[1]は密度を基にしたクラスタリングで，任意の形状のクラスターを発見できるとともに，ノイズを検出することも可能である．一方で，*EPS* と *MinPts* という 2 つのパラメータを入力する必要があり，クラスタリング結果は分析者に依存する．本研究では，この DBSCAN におけるパラメータ設定を自動化し，分析者に依存せず，クラスタリング結果を出すことが可能な手法を提案する．

2. 提案手法

本研究では，先行研究とし Auto eps DBSCAN[2]を採用し，この手法を比較対象として用いる．DBSCAN をベースとした Auto eps DBSCAN は，パラメータ k を入力することで，k-dist plot と呼ばれる密度分布を把握するグラフを作成し，そこから *EPS* と *MinPts* を自動推定し，クラスタリングを行う手法である．k を入力する目的はデータの分布から異なる密度領域とノイズ部分を判断するためであり，k が小さい値の場合，ノイズがクラスターとなる密度となってしまうため，適さない．一方で k が大きい値の場合，密度の違いが判断できなくなってしまう．そこで，k を増加させたときの密度分布のグラフの値の変化が小さくなった段階の k を選択することにより，k を自動決定するアルゴリズムを提案する．加えて，k-dist plot の密度の変化およびノイズ部分の検出方法，*MinPts* の計算方法の改善も合わせて行った．

3. 実験結果

今回は 3 種類の人工的に生成したデータセットに対して，既存手法と提案手法の比較実験を行った．既存手法は k の値を 1~50 の間で変化させ，それぞれの k に対して評価指標を計算し，それらの最適値を既存手法の結果とすることにした．その結果を以下の表 1 に示す．これより，ほぼすべての指標に関して提案手法は既存手法を上回り，提案手法が有効であることが示された．

表 1　比較実験結果

データ	手法	k	シルエット係数	RI	ARI	正規化相互情報量(NMI)	調整正規化相互情報量(AMI)	purity	entropy
データセット1	既存手法の最適値		**0.552**	0.906	0.883	0.897	0.892	0.948	0.105
	提案手法	12	0.487	**0.945**	**0.932**	**0.923**	**0.921**	**0.972**	**0.077**
データセット2	既存手法の最適値		0.193	0.698	0.633	0.816	0.696	0.786	0.280
	提案手法	13	**0.371**	**0.821**	**0.794**	**0.850**	**0.829**	**0.884**	**0.125**
データセット3	既存手法の最適値		**0.495**	0.927	0.902	0.886	0.863	0.910	0.117
	提案手法	11	0.061	**0.969**	**0.958**	**0.933**	**0.905**	**0.957**	**0.036**

参考文献

[1] M. Ester, et al. (1996). “A density-based algorithm for discovering clusters in large spatial databases with noise“. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining(KDD-96), 226-231
[2] M. N. Gaonkar and K. Sawant. (2013)."AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset". International Journal on Advanced Computer Theory and Engineering, 2(2), 11-16