5N-2

データマイニングにおけるデータ精度の精錬に関する研究

権　相鑫†

東京電機大学大学院情報環境学研究科†

1．はじめに

近年，大容量記憶媒体の低価格化，計算機処理能力の向上，情報通信技術の急速な進展の効果があいまって，ネットワーク社会におけるデータの収集や活用が近年格段に容易になった．実際，計算機で処理できる多くの情報がインターネットを通じて世界中を飛び交い，これらの情報に誰もがいつでも自由にアクセスできる時代に突入した．データマイニングは時代の要請に応えるべく生まれてきた技術である[1]．クラスタリングはデータマイニングの重要なツールとして利用され[2]，大規模データ処理などの新たな要求が生じている．近年，これらの要求に対処する様々な手法が研究されるようになっている．

大規模データには不審なデータを含まれている．そして不審なデータを発見，排除して設定した個数のグループを分割する研究が必要である．

本研究は二つのクラスタリング本研究は大規模データを扱う２つのクラスタリングアルゴリズムを採用している．DBSCAN (Density-Based Spatial Clustering of Applications with Noise)密度ベースの空間クラスタリングとK-means （K-平均法）クラスタ数K個に分類する非階層型クラスタリングである．二つのクラスタリングを扱って大規模データの中に不審なデータを発見，排除した後事前に設定したクラスタ数によってクラスタリングの結果を求める．そしてデータ精度を精錬することを目標とする．

2. 既存研究

クラスタリングはさまざまな手法が提案されているが，大きく分けるとデータの分類が階層的になされる階層型手法と，特定のクラスタ数に分類する非階層的手法とがある．他には，密度ベースのクラスタリング方法もある．

２.１　K-means クラスタリング

K-meansクラスタリングは非階層型クラスタリング手法である．入力されたデータをK個に分ける．最初は適度にK個の点を取り，その点に近いもの同士が同じクラスタになるように，データを分割する．

Study on the refining of data accuracy in data mining
†Xiangxin Quan
†Graduate School of Information Environment, Graduate School of Tokyo Denki University

K-means クラスタリングは K 値によって全データを分割するが，不審なデータを発見することができない．

２．２　DBSCAN クラスタリング

DBSCAN クラスタリングは密度ベースの空間クラスタリングという手法である．DBSCAN クラスタリングは距離のしきい値と対象数のしきい値という二つのパラメータを用いる[3]．全データの中で低密度領域から分離した高密度領域を探す．そして任意形状のクラスタを処理することができる．

DBSCAN クラスタリングは自動的にクラスタ数を生成してデータを分割する．事前に設定したクラスタ数によって分割することができない．

3．提案方法

本研究はデータの中に不審なデータを発見，排除した後事前に設定したクラスタ数によってクラスタリングの結果を求めるために，K-means クラスタリングと DBSCAN クラスタリングを結合して研究する．提案手法の流れは図１ように示す．

1. DBSCAN クラスタリングの距離のしきい値と対象数のしきい値、K-means クラスタリングの K 値を設定する．
2. DBSCAN クラスタリングを使ってデータ中の不審なデータを発見，排除する．
3. K-means クラスタリングを使って不審データを排除したデータをもう一回クラスタリングする．
4. 設定したクラスタの数によってクラスタリングの結果を求める．

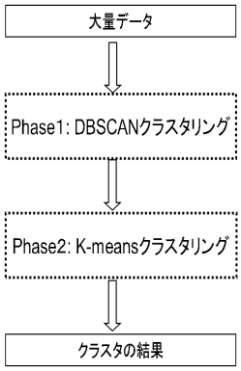

図. 1 提案手法の流れ

4. 実験

4.1 実験目的

実験に用いたデータは KDD-cup-99 第三回国際知識発見とデータマイニングツールコンクールに使用されているデータ・セット[4]を使った．本実験はデータ・セットの中のネットワークの接続のデータバイト数のデータ(500 件)を利用して実験をする．

実験には K 値を 5，距離のしきい値を 5，対象数のしきい値を 5 に設定した。そしてネットワークの接続のデータバイト数が大きい接続とデータバイト数が少ない接続を不審の接続と設定した．

実験の目的は不審のデータがない接続を五つのクラスタに分割して求める。最後に，データ精度を精錬することを求める．

4.2 実験の流れ

提案した DBSCAN and K-means クラスタリングの流れは下のように示す．

- まず，DBSCAN クラスタリングを利用してデータ中の不審のデータを排除する．
- 次に，K-means クラスタリングを利用して不審のデータを排除したデータをクラスタリングする．
- 最後に，クラスタの個数を決めたクラスタリングの結果を求める．

最後に，K-means クラスタリングの結果，DBSCAN クラスタリングの結果，DBSCAN and K-means クラスタリングの結果を比べる．

4.3 実験結果

表.1 K-means の結果

クラスタ	接続の数
1	68
2	173
3	36
4	54
5	169

表.2 DBSCAN の結果

クラスタ	接続の数
1	5
2	364
3	61
4	9
5	55
6	6

表.3 DBSCAN and K-means の結果

クラスタ	接続の数
1	64
2	218
3	138
4	54
5	11

ネットワークの接続のデータバイト数のデータ(500 件)を K-means クラスタリング実験の結果は表 1 のように五つのクラスタが出た．

DBSCAN クラスタリングの結果は表 2 のように六つのクラスタが出た．そして，クラスタ 1，4，6 の接続の数が少ない．

提案した DBSCAN and K-means クラスタリングの結果は表 3 のように五つの結果が出た．本結果は異常な接続(データバイト数が大きい接続とデータバイト数が少ない接続)20 件を排除した後，五つのクラスタの結果が出た．

4.4 実験の検証

エントロピーは最も標準的に用いられている評価尺度．値が低いほどクラスタリング結果がよい．E_iはクラスタ i のエントロピー、p_iは生起確率、N はデータの数である．精度の比較は表.4 のように示す．

$$E_i = -\sum p_i \log_2 p_i$$

$$\text{平均情報量} = \sum E_i / N$$

表.4 精度の比較

精度評価	平均情報量
k-means	0.002868
DBSCAN	0.001798
DBSCAN and k-means	0.001787

5. おわりに

今回は，DBSCAN クラスタリングと K-means クラスタリングの問題点，それに対する提案方法である．そして，実験用データを見つけ，実験の流れを．後に実験を行い，クラスタの結果を取った．

本研究に提案した DBSCAN and K-means クラスタリングのエントロピーと DBSCAN クラスタリングのエントロピーはほぼ同じだった．

本研究に提案した DBSCAN and K-means クラスタリングはデータの中で、異常なデータのクラスタを発見し、決めている個数のクラスタを求めることができた．

今後は大量のデータの中でクラスタリングした結果の評価について検討していきたい．

参考文献

[1] 元田 浩， 津本 周作， 山口 高平， 沼尾 正行， “データマイニングの基礎 ” 情報処理学会， 編集， pp. 1-１１， Aug 200

[2] ローネン・フェルドマン， ジェイムズ・サンガー， “テキストマイニングハンドブック” IBM 東京基礎研究所翻訳， pp. 111-116, Sep 2010

[3] 神嶌敏弘， “データマイニング分野のクラスタリング手法” 産業技術総合研究所， 情報処理学会， pp. 174-176, Sep 2007.

[4] KDD-cup-99 第三回国際知識発見とデータマイニングツールコンクールにデータ・セット http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html