

クラスタリングとNNの組み合わせアルゴリズムに基づくデータ分類

YANG, Yingfei / 楊, 瑩菲

(出版者 / Publisher)

法政大学大学院理工学研究科

(雑誌名 / Journal or Publication Title)

法政大学大学院紀要. 理工学研究科編

(巻 / Volume)

64

(開始ページ / Start Page)

1

(終了ページ / End Page)

2

(発行年 / Year)

2023-03-24

(URL)

<https://doi.org/10.15002/00026414>

クラスタリングと NN の組み合わせアルゴリズムに基づく データ分類

Classification of Datasets Based on Combination Algorithm of Clustering and Neural Network

楊瑩菲

Yingfei YANG

指導教員 李 磊

法政大学大学院理工学研究科応用情報工学科専攻博士前期課程

With the development of the data era, data's importance is increasing daily. For data analysis, the accuracy of data classification is again the basis to ensure the data analysis is carried out smoothly. Therefore, in order to improve the accuracy of the classification of unknown data, this paper proposes a new approach based on combinatorial clustering and neural network algorithm. The results show that the algorithm combining clustering algorithm and neural network will help to improve the accuracy of data classification and identification effectively.

Key Words : hierarchical clustering; k-means; neural network; data classification

1. はじめに

データ時代な発展に伴い、データの重要性はますます高まっており、データマイニングも急速に発展している。クラスタリング分析はデータ分析の重要な分野であり、社会発展におけるデータ精度要件が継続的に言及されているため、アルゴリズムのクラスタリング精度を向上させる方法は人気のある研究テーマである。

先行研究は最初の分類は直接的に非階層クラスタリング k-means を用いた。本研究では、階層クラスタリングから K 値を求めて、BP ニューラルネットワークに基づくクラスタリングを提案する。

2. アルゴリズム

(1) クラスタリングアルゴリズム

クラスタリング (clustering) とは、機械学習における教師なし学習の 1 種で、データ間の類似度にもとづいて、データをグループ分けする手法である。

クラスタリングの計算方法を大きく分類すると、階層構造を作りながらグルーピングを進める「階層クラスタリング」と、階層構造を作らずに分類していく「非階層クラスタリング」の 2 つに分けられる。階層クラスタリングと非階層クラスタリングにはそれぞれ特徴があり、それぞれの代表的な手法として Ward 法(階層)、K-means (非階層)法などがある。

エルボー法は、階層型クラスタリングを行う際に、適切なクラスタ数を決定するための一般的な方法である。エルボー法は、各クラスタ内のばらつきの指標であるクラ

スタの重心におけるクラスタ内二乗距離の和のグラフにおいて、最大の「エルボー」を生み出すクラスタ数が最適であるという仮定に基づくものである。最適なクラスタ数は、異なるクラスタ内二乗の数を比較することによって決定される。

(2) k-means 法

k-means 法は、データを適切なクラスタに分割し、そのクラスタの平均値を用いてデータを調整するアルゴリズムである。また、k-means 法は、任意に指定した k 個のクラスタを作成するアルゴリズムであるため、k 点平均法とも呼ばれている。

k-means 法の具体的な手順を下記で示す。① 各点 x_i に対してランダムにクラスタを割り振る。

② 各クラスタに割り当てられた点について重心を計算する。

③ 各点について上記で計算された重心からの距離を計算し、距離が一番近いクラスタに割り当て直す。

④ step②と③を、割り当てられるクラスタが変化しなくなるまで行う。

(3) 正規化(Z-score)

Z 得点 (z 値、z-score、z-value) とは、平均が 0、標準偏差 (SD) が 1 になるように変換した得点。

母集団の平均と標準偏差が既知の場合、変数 x の Z 得点 z は以下のように計算される。

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

ここで、 μ は母集団の平均、 σ は母集団の標準偏差を

表す。

3. 提案されるアルゴリズム

(1) K-means&NN 組み合わせアルゴリズム

先行研究は最初の分類は直接的に非階層クラスタリング k-means を用いた。本研究では、階層クラスタリングアルゴリズムから、分類数 K を確認し、k-means 法を利用して、k-means の結果はニューラルネットワークの訓練データとし、分類するアルゴリズムを提案する。

(2) 手順

① 全データの中から一部のデータをランダムに選び、グループ A とする。グループ A のデータは Z スコアを用いて標準化する。

② 標準化されたグループ A のデータについて、階層型クラスタリングアルゴリズムを用いて k 値を求める(データはいくつかのカテゴリに分けることができる)。

③ 得られた k 値を k-means の k として入力し、データをグループ A に分類する。

④ グループ A のデータをニューラルネットワークの入力とし、グループ A のデータの分類結果をニューラルネットワークの出力とし、ニューラルネットワークを学習させて分類モデルを生成する。

⑤ 生成したモデルを全データに適用し、分類結果を得る。

(3) フローチャート

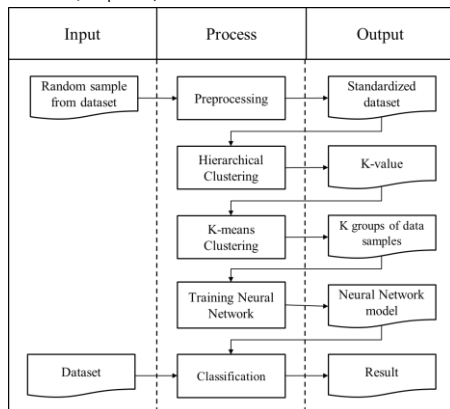


図1 提案するアルゴリズムのフローチャート

4. 実験

(1) 実験データ

今回の実験はあやめ、麦の種、携帯のデータ三種類のデータを利用している。

表1 実験データの比較

	Number of categories	Number of variables	Data Size
Iris	3	4	150
Seed	3	7	210
Phone rank	4	14	600

(2) 実験結果

今回の実験、k-means と提案するアルゴリズム実行結果は次の表2と図2を示す。

表2 正確率の比較

	K-means	K+NN
Iris	89.33%	96.67%
Seed	89.52%	92.86%
phone	42.21%	66.33%

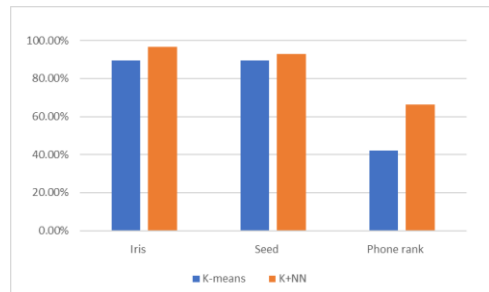


図2 正確率の比較

5. まとめ

いままで、k-means と改良したアルゴリズムを用いて、あやめ、麦の種、携帯のデータで実行しました。分類結果の正確率が上がったことを知った。

今後では、ランダムではなく、どうやって特徴なデータを選択するについて、研究を続けていくと思う。

参考文献

- 1) Z. Muda, W. Yassin, M. N. Sulaiman and N. I. Udzir. "Intrusion detection based on K-Means clustering and Naïve Bayes classification." 2011 7th international conference on information technology in Asia. IEEE, 2011.
- 2) Liu Yingqiu, Wei Li, and Li Yunchun. "Network traffic classification using k-means clustering." Second international multi-symposiums on computer and computational sciences (IMSCCS 2007). IEEE, 2007.
- 3) Zhu, Changsheng, Christian Uwa Idemudia, and Wenfang Feng. "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques." Informatics in Medicine Unlocked 17 (2019): 100179.
- 4) Trupti M. Kodinariya, and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." International Journal 1.6 (2013): 90-95.
- 5) Ahmed M. Abdel-Zaher, and Ayman M. Eldeib. "Breast cancer classification using deep belief networks." Expert Systems with Applications 46 (2016): 139-144.
- 6) Omar Ahmed, and Adnan Brifcani. "Gene expression classification based on deep learning." 2019 4th Scientific International Conference Najaf (SICN). IEEE, 2019.
- 7) Amjad Rehman Khan, et al. "Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification." Microscopy Research and Technique 84.7 (2021): 1389-1399.