

Naruaki TOMA <tnal@ie.u-ryukyu.ac.jp>
宛先: "slab34-dm@ie.u-ryukyu.ac.jp" <slab34-dm@ie.u-ryukyu.ac.jp>
返信先: slab34-dm@ie.u-ryukyu.ac.jp
[slab34-dm:26] Level 6: クラスタリングについての検討

2013年5月30日 17:14

=====
Level 6: クラスタリングについての検討

べ切: 6/6, 次回実験にて報告してもらいます
提出方法: Wiki

機械学習の代表的な使い道の一つである「クラスタリング」では、
「サンプル (N次元の特徴ベクトル) 集合」として用意された
データセットにおいて、似たようなサンプル同士を「同じクラスタ」
に分類することでサンプル全体をいずれかのクラスタに分類する
ことを目的としている。

手法の特徴としてはK-meansに代表されるトップダウンな手法、
ワード法に代表されるボトムアップ的手法に大別することが多い。
どの手法も一般的には、「サンプル間の距離をどのように計測するか」、
「初期値をどのように設定するか」等によってクラスタリング
結果が大きく異なることも珍しくない。そのため、適切なクラスタ
を得るために、
「適切なクラスタ ≒ 良く見られる (現れる) クラスタ ≒ 意味がありそうなクラスタ」
と考え、複数回試行した結果を俯瞰して観察することも少なくない。

上記のような特性を踏まえた上で、クラスタリングについて
以下の点について検討し、報告せよ。

(1) 適用してみたい具体的なクラスタリング問題について述べよ。
どのようなデータセットにクラスタリングを適用したい？
各サンプルの特徴ベクトルは2次元以上とする。
可能な範囲でサンプルを具体的な特徴ベクトルで複数事例表現してみよう。
またそのデータセットにクラスタリングを適用する意図は何だろうか？

(2) 一般的にクラスタリングには教師データが無いことが多い。
教師データが揃うなら教師あり学習 classification した方が良い。
クラスタリング結果の妥当性や適切さについて、どのように評価したら良いだろうか？

Naruaki Toma
E-mail: tnal@ie.u-ryukyu.ac.jp, Tel: 098-895-8830
<http://www.eva.ie.u-ryukyu.ac.jp/~tnal/>

--
You received this message because you are subscribed to the Google Groups "slab34-dm" group.
To unsubscribe from this group and stop receiving emails from it, send an email to slab34-dm+unsubscribe@ie.u-ryukyu.ac.jp.
To post to this group, send email to slab34-dm@ie.u-ryukyu.ac.jp.
Visit this group at <http://groups.google.com/a/ie.u-ryukyu.ac.jp/group/slab34-dm/?hl=en-US>.