Recent Developments in Document Clustering

4. Extensions to *k*-means

茨城大学工学部情報工学科
08T4018Y　小幡智裕

階層的クラスタリング手法

- 各文書は、はじめのうちはそれ自身がクラスタ
- 連続的に文書を合併し、階層構造で表わす
- 階層は樹形図や木とも呼ばれ、階層の根は、コーパス内のすべての文書を含む。

分割クラスタリング手法

- 階層的アルゴリズムよりも優れているといわれている。

- 階層構造が強く望まれる場合、階層構造の各レベルで分割手法を用いることでより優れた結果を得ることができた。

k-means(k平均法)

- *k*-meansは古典的な分割手法。
- 文書を初期クラスタとしてk個に分割することから始まり、これらのクラスタのセントロイドを計算し、繰り返しドキュメントをクラスタに割り当てる。

k-meansの欠点

1. 一般的に初期化する量に依存する。

2. ノイズや異常値が連続する。

3. 複雑さがO(nkl)

球面*k*-means

- テキストデータには方向性があることが知られている。
  →ユークリッド距離より、コサイン類似度を用いて文書のベクトルを標準化する。

- このようなアルゴリズムを球面*k*-meansという。

4.1 Online spherical *k*-means

- 球面*k*-meansを拡張したもの。クラスタリングの速度、類似度、正確さを向上させるため、競合学習技術を用いている。

- クラスタリングとして文書を追加する際、学習率によって処理される。

学習率

・目的関数

$$J = \sum_{d \in D} d^T \mu_{k(d)}$$

オンライン球面k-meansも球面k-meansもこの式を最小化することを目的とする。dはコーパスD内の文書ベクトルを表わす。そして、$k(d) = \arg\max_k d^T \mu_k$ は、クラスタの中心を表わす。$d^T \mu_{k(d)}$ は、文書ベクトルとセントロイドの内積を表わしている。

学習率（続き）

・セントロイドではなく、学習率ηを導入する。

$$\mu'_{k(x)} = \frac{\mu_{k(d)} + \eta d}{|\mu_{k(d)} + \eta d|}$$

学習率ηは今後の入力に文書が適応するかを制御する。

最もよく働く学習率は、

$$\eta_t = \eta_0 \left(\frac{\eta_f}{\eta_0} \right)^{\frac{t}{NM}}$$

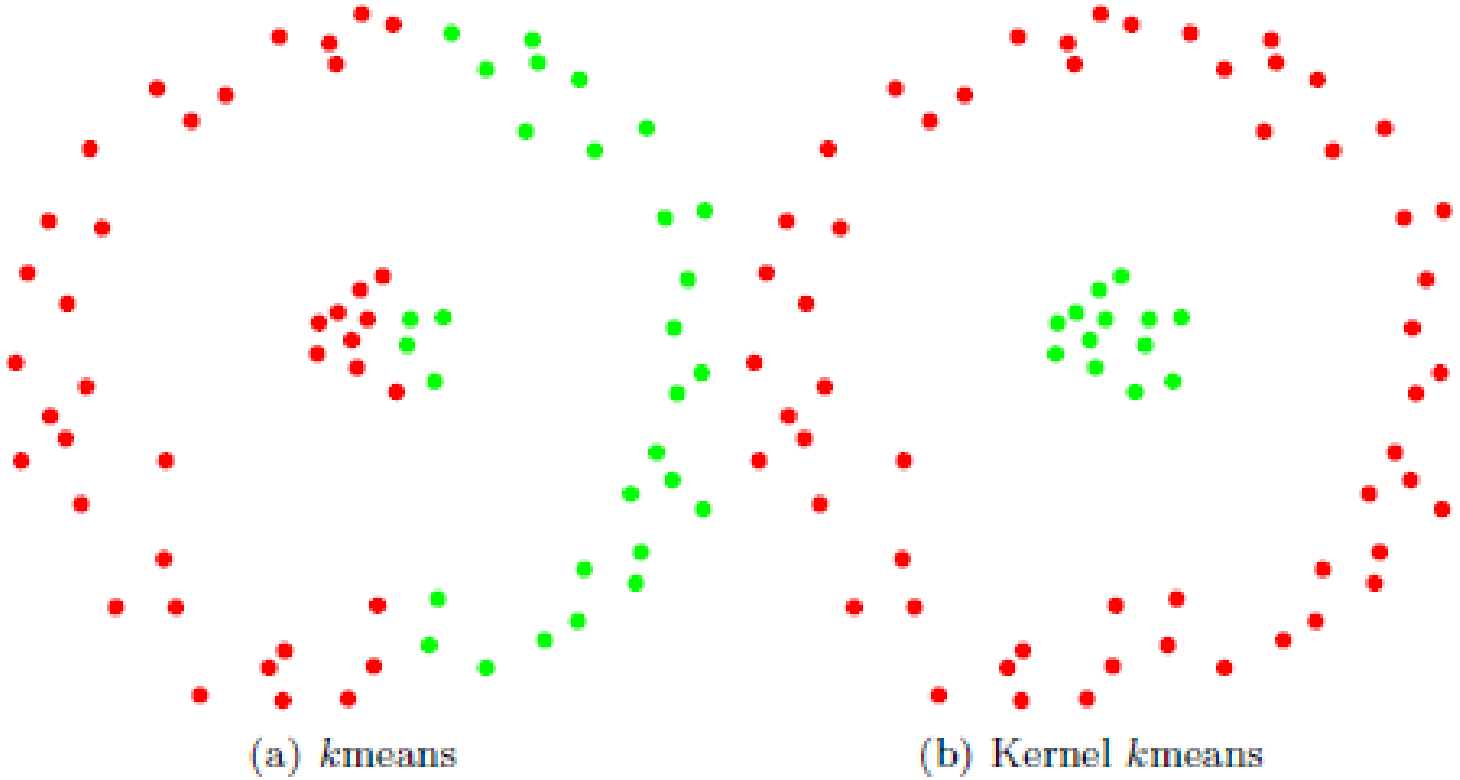
である。Nはドキュメント数、Mはバッチの反復回数、ηfは最終学習率(ηf =0.01)である。

メリット

- 文書がコーパスに追加されるとき、文書の全体集合にバッチクラスタリングを返す必要がない。

- オンライン球面k-meansは、球面の2倍速く収束し、リアルタイムのアプリケーションに適しているといえる。

4.2 kernel *k*-means



(a) *k*means (b) Kernel *k*means

Kernel関数

・次のように定義される

$$\kappa(d_i, d_j) = \tanh(c(d_i, d_j) + \theta)$$

これは、文書の組の類似性を測定するものである。

カーネル関数kはn×nの正方行列Kで表わすことができる。

この公式は、各文書の重みづけとして割り当てる。

diagonal dominence

- 文書と他の文書がよく似ているとみなされた時に起こる。

- 結果として得られた類似度行列は値がとても大きい対角優位行列になる。

メリット

- kernel k-meansはマルチステージアルゴリズム（セクション7.3）よりも優れていると証明された

- 教師なしのカーネル法は、オーバーラッピングクラスタに適用することができる。