テキストデータに対するエージェンティッククラスタリングの検証

林和希 [1]　参木裕之 [1]
[1] 株式会社大和総研 デジタルソリューション研究開発部
{kazuki.hayashi,hiroyuki.mitsugi}@dir.co.jp

概要

テキストクラスタリングは、クラスタ数の事前不明性や短文の情報量制約により依然として困難な課題である。本研究では、エージェンテックなアプローチとして、大規模言語モデル（LLM）を用いてクラスタ数を推定し、その結果を K-means や階層クラスタリングなどのアルゴリズムに適用する新たな手法を提案する。提案手法は、トピック分類タスクにおいて従来手法（エルボー法、BERTopic 等）を上回る精度（ARI、NMI）を示した。一方、感情分類など極性情報を含むデータセットでは、プロンプト設計や情報表現の工夫が今後の課題として残された。

1　背景

テキストなどの高次元データから、内在する意味のあるグループ（クラスタ）を探索することは、データマイニング分野における重要な課題であり、名寄せなど産業応用の面からも需要が大きい。

この課題に対応するため、古典的な K-means 法の改良型手法 [1]、BERTopic に代表されるトピックモデリング手法 [2]、さらには DEC[3] などの深層クラスタリング手法が活発に研究されてきた。一方で、これらの手法には共通してクラスタ数 K の決定が自明ではないという課題が存在する [4]。階層的クラスタリングや HDBSCAN（BERTopic で利用される手法）のようにクラスタ数を直接指定しない手法もあるが、これらも別のパラメータ設定を必要とし、その適切な設定は容易ではない。

クラスタ数やパラメータを決定するために、エルボー法 [5] や内部評価指標（Silhouette 係数、Davies–Bouldin 指数、Calinski–Harabasz 指数など）が用いられる。しかし、これらはクラスタ形状に依存するバイアスを含み、必ずしも適切な結果を保証しない [6]。さらに、トピックモデリングではコヒーレンススコアを最適化する手法があるが、クラスタ数を増やすとスコアが上昇する傾向があり、過分割を誘発する問題が指摘されている [7]。このように、アルゴリズムや数理最適化のみで適切なクラスタリング結果を得ることは困難である。

近年、この課題に対応する新たな潮流として、大規模言語モデル（LLM）をクラスタリングに応用する研究が注目されている。Viswanathan らは、LLM の few-shot 学習能力を活用し、半教師ありクラスタリングを実現する手法を提案した [8]。しかし、このような手法は人間による few-shot 例の提示を前提としており、データセットが大規模な場合、適切な例を選定することは現実的に困難である。

本研究では、人手による例示を必要としない LLM 活用の可能性を探る。具体的には、AI エージェントによるテキストクラスタリング（埋め込みベクトル表現の作成、次元削減、アルゴリズム選択、パラメータ設定など一連の工程）の自動化を最終目標とし、その第一歩としてクラスタ数推定に LLM を活用する手法を提案する。

2　提案手法

図 1 に、上段に提案手法、下段に最終的な理想形のフローを示す。

提案手法は以下の 4 ステップで構成される。

1. **埋め込み表現への変換**
   テキストデータセットを gemini-embedding-001 で 3,072 次元の L2 正規化済みベクトルに変換 (図 1 ※ 1)。
2. **次元削減と標準化**
   UMAP(パラメータは BERTopic デフォルト設定) で次元削減後、z-score 標準化（図 1 ※ 2,3)。
3. **LLM によるクラスタ数推定**
   次元削減前のベクトル間の内積に基づき、意味的に近い要素が隣接するよう並べ替えて、LLM

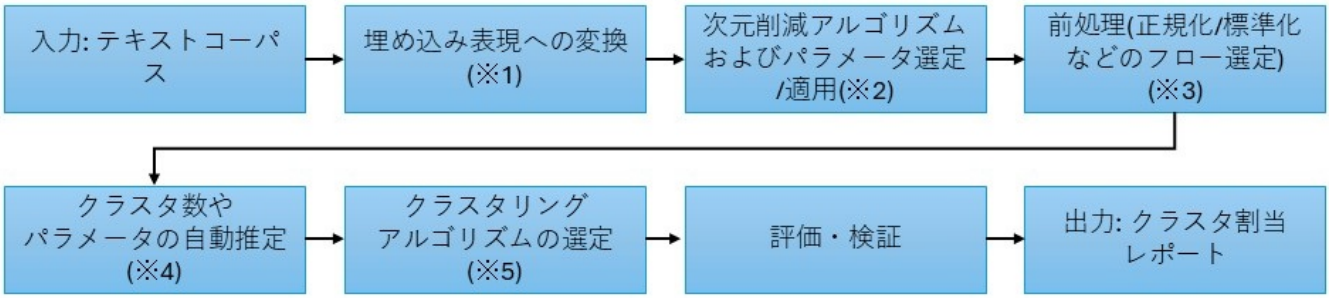

図 1: エージェンテック・クラスタリングのフロー

表 1: 使用したデータセット一覧

(シルエットスコアはサブデータセット 3.2.2 の次元削減後埋め込み表現とラベルから計算)

言語	データセット名	サンプル数	クラスタ数	シルエットスコア
英語	20newsgroup	4531	20	0.227
英語	ag-news	120,000	4	0.301
英語	bbc	2,225	5	0.538
英語	yahoo	60,000	10	0.113
日本語	livedoor news jmteb	1,106	9	0.254
日本語	massive scenario classification jmteb	11,514	18	0.309
日本語	mewsc16 ja jmteb	992	12	0.134
日本語	sib200 japanese classification jmteb	701	7	0.183
日本語	健康経営度調査 (令和 3 年度)	3,485	10	0.417
日本語	健康経営度調査 (令和 4 年度)	3,954	10	0.392
日本語	健康経営度調査 (令和 5 年度)	4,276	10	0.408
日本語	健康経営度調査 (令和 6 年度)	4,602	11	0.325

がクラスタ構造を把握しやすくソートしたテキストを gemini-2.5-flash に入力し、Structured Output でクラスタ数 K を推定。再現性確保のため、thinking budget=0、温度=0、シード値=42 を設定（図 1 ※ 4）。

4. **クラスタリング**
   推定された K を用いて、K-means 法および階層クラスタリングを適用（図 1 ※ 5）。

本手法はクラスタ数推定の精度向上と再現性を重視し、将来的にはクラスタラベル付与や評価指標の自動生成まで含む完全自動化を目指す。

3 実験

テキストクラスタリングの応用領域は多岐にわたるが、本研究では代表的な問題設定としてトピック分類に着目し、これらのデータセットを対象に手法の有効性を検証した。

3.1 評価指標

クラスタリング結果の品質を評価するため、以下の指標を用いた。

3.1.1 クラスタ数推定誤差

各データセットにはそれぞれテキストとそれに対応するクラスラベルが付与されている。この際、ユニークなクラスラベル数のことを「真のクラスタ数」と呼称する。したがって、LLM が推定したクラスタが「真のクラスタ数」と比べてどれほど離れているのかを表す推定誤差 (真のクラスタ数 – *LLM* が推定したクラスタ数) を評価指標とした。

3.1.2 Adjusted Rand Index (ARI)

ARI は、ランダム一致の影響を除去した上でクラスタリング結果の一致度を評価する指標であり、次式で定義される：

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}} \quad (1)$$

ここで、クラスタリング結果を U と V 、n_{ij} はクラスタ U_i とクラスタ V_j に共通して含まれる要素数を表す。また $a_i = \sum_j n_{ij}, b_j = \sum_i n_{ij}$ であり、n は全要素数を表す。

3.1.3 Normalized Mutual Information (NMI)

NMI は、クラスタリング結果間の情報共有度を測る指標であり、次式で定義される：

$$NMI(U,V) = \frac{2 \cdot I(U;V)}{H(U) + H(V)} \tag{2}$$

ここで、$I(U;V)$ は相互情報量、$H(U), H(V)$ はクラスタリング U および V のエントロピーを意味する。

3.2 実験設定

3.2.1 手法

本研究では、提案手法として「LLM によるクラスタ数推定を用いた K-means クラスタリング（**提案手法 1**）」および「LLM によるクラスタ数推定を用いた階層クラスタリング（**提案手法 2**）」を採用した。比較手法としては、「エルボー法による K-means クラスタリング（**従来手法 1**）」、「エルボー法による階層クラスタリング（**従来手法 2**）」、ならびに標準パラメータ設定の BERTopic を用いた。

3.2.2 データセット

表 1 に使用したデータセットを記す。健康経営度調査とは評価結果の開示に同意した健康経営度調査回答法人に対する評価結果（フィードバックシート）をまとめたものであり、各社の課題内容テキストおよび課題テーマ番号 (ラベル) が記載されている。なお、[9] では施策のタグ付けのためクラスタリングを実施している。

今回の実験では、これらのデータセットから 500 件、1,000 件、1,500 件、2,000 件の 4 パターンで層化サンプリングを行い、1 つのデータセットから 4 つのサブデータセットを生成した。

更に、それぞれのサブデータセットに対して各手法でクラスタリングを実施し、各手法の性能を検証した。なお、データ数が 1,000 件未満の場合はサブデータセットは 500 件のもののみを生成した。

3.3 結果

3.3.1 クラスタ数推定精度

図 3 にトピック分類の各サブデータセットに対してクラスタリングを実施した際のクラスタ数の推定誤差をプロットした。図 3 に示すように、LLM は推定誤差が少ない傾向がある一方で、BERTopic はクラスタ数を過大に、エルボー法は過少に見積ってしまう傾向があった。このことから、特にトピック分類においては LLM によるクラスタ数の推定が有用であると考えられる。

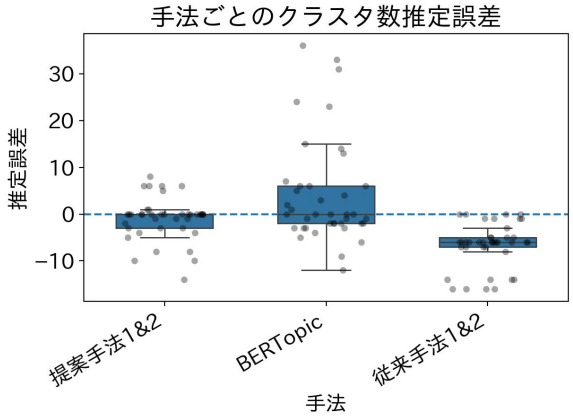

図 2: クラスタ数推定精度

3.3.2 Adjusted Rand Index (ARI)

各手法の ARI を表 2 に示す。なお、サブデータセットが複数存在する場合、それらの平均値を各データセットの評価の平均値とした。結果として、AG-NEWS と sib200 を除く全てのデータセットで提案手法が従来手法を上回った。また、多くのケースで BERTopic を超える性能を示した。特に、健康経営度調査や massive scenario classification jmteb では顕著な改善が見られる。総合的な評価としては、提案手法 1 が最も高い ARI となった。

3.3.3 Normalized Mutual Information (NMI)

各手法の NMI を表 2 に示す。なお、サブデータセットが複数存在する場合、それらの平均値を各データセットの評価の平均値とした。結果として、提案手法はほとんどのデータセットで従来手法を上回り、特に massive scenario classification jmteb や mewsc16 ja jmteb で顕著な改善が見られた。一方、ag-news では従来手法が優位であった。健康経営度調査では BERTopic が最も高い NMI を示したが、提案手法も競合する性能を示した。総合的な評価としては、提案手法 1 が最も高い NMI となった。

4 議論

本研究では、AI エージェントによるテキストクラスタリングの自動化に向け、LLM を用いたクラスタ数 K の推定手法を提案し、従来手法（エルボー法、BERTopic 等）と比較した。

表 2: 各手法およびデータセットの ARI と NMI 一覧

データセット名	ARI					NMI				
	提案手法 1	提案手法 2	従来手法 1	従来手法 2	BERTopic	提案手法 1	提案手法 2	従来手法 1	従来手法 2	BERTopic
20newsgroup	**0.451**	0.444	0.207	0.198	0.322	**0.659**	0.651	0.495	0.490	0.639
ag-news	0.379	0.354	0.532	**0.533**	0.166	0.480	0.479	**0.530**	0.519	0.371
bbc	0.657	**0.669**	0.568	0.565	0.422	0.755	**0.777**	0.706	0.710	0.617
yahoo	**0.356**	0.343	0.206	0.187	0.117	**0.448**	0.444	0.320	0.321	0.341
livedoor news jmteb	0.526	**0.569**	0.413	0.417	0.441	0.649	**0.663**	0.594	0.603	0.605
massive scenario classification jmteb	**0.515**	0.499	0.203	0.224	0.402	**0.723**	0.717	0.476	0.499	0.717
mewsc16 ja jmteb	**0.416**	0.406	0.344	0.376	0.273	**0.575**	0.548	0.398	0.483	0.498
sib200 japanese classification jmteb	0.317	0.317	**0.437**	0.413	0.330	**0.523**	0.505	0.468	0.467	0.470
健康経営度調査 (令和 3 年度)	0.641	**0.665**	0.413	0.393	0.639	0.737	0.748	0.652	0.630	**0.760**
健康経営度調査 (令和 4 年度)	0.587	0.600	0.424	0.408	**0.614**	0.744	0.745	0.650	0.639	**0.757**
健康経営度調査 (令和 5 年度)	**0.653**	0.622	0.446	0.405	0.644	0.764	0.757	0.671	0.646	**0.786**
健康経営度調査 (令和 6 年度)	0.577	0.563	0.416	0.424	**0.621**	0.726	0.735	0.630	0.630	**0.779**
平均値	**0.506**	0.504	0.384	0.378	0.416	**0.649**	0.647	0.549	0.553	0.612

4.1 実験結果の考察

実験の結果、特にクラスタ数が多い massive scenario classification jmteb や健康経営度調査データセット（令和 3 年度、令和 5 年度）において、提案手法は ARI・NMI ともに従来手法を大きく上回った。

一方、AG-NEWS や健康経営度調査データセット（令和 4 年度、令和 6 年度）では、一般的な分類基準と異なる粒度でラベル付けされているため、LLM によるクラスタ数推定が困難であった。

また、提案手法は K-means を使用したが、不均衡なデータに対しては K-means よりも密度ベースの BERTopic（HDBSCAN）が高い性能を示す傾向があるため、今後は密度ベース手法への適用も検討する必要がある。

4.2 シルエットスコアとの関係

表 1 のシルエットスコアは、次元削減後の埋め込み表現 X とラベル y から計算した値であり、この値が低いほどクラスタリングが難しいことを示す。特に Yahoo データセットではシルエットスコアが低いにも関わらず、提案手法は ARI・NMI ともに従来手法を大きく上回った。このことから、明確な区分が難しいデータに対しても提案手法の有効性が示された。

4.3 従来手法との比較

エルボー法など従来のクラスタ数推定手法は、データ分布やクラスタ形状に強く依存し、高次元・非凸構造では精度が低下しやすい。一方、LLM は言語的意味構造を反映できるため、数値的ヒューリスティックでは困難なクラスタ数推定が可能となった点は大きな意義である。

4.4 課題

提案手法には以下の課題が残る。

- LLM への入力として全テキストを渡す必要があり、コンテキスト長制約から大規模データへの適用が難しい。
- クラスタ数を明示的に用いない密度ベース手法（例：DBSCAN 等）への拡張には追加的な工夫が必要である。
- 感情分類のような極性情報を含むタスクでは、プロンプト設計や埋め込み表現の工夫が不可欠である。

今後は、プロンプト設計の最適化、部分的サンプリングや要約によるスケーラビリティ向上、極性情報の明示的活用、多言語・大規模データへの適用、密度ベース手法への拡張などを検討する。

5 結論

本研究では、LLM を用いたクラスタ数推定手法を提案し、従来手法と比較した。トピック分類タスクにおいては、提案手法がエルボー法や BERTopic を上回る性能を示した。今後は、プロンプト設計やスケーラビリティの向上、密度ベース手法への拡張などを進める予定である。

参考文献

[1] Marieke E Timmerman, Eva Ceulemans, Kim De Roover, and Karla Van Leeuwen. Subspace k-means clustering. **Behavior research methods**, Vol. 45, No. 4, pp. 1011–1023, 2013.

[2] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. **arXiv preprint arXiv:2203.05794**, 2022.

[3] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In **International conference on machine learning**, pp. 478–487. PMLR, 2016.

[4] Collin Leiber, Niklas Strauß, Matthias Schubert, and Thomas Seidl. Dying clusters is all you need-deep clustering with an unknown number of clusters. In **2024 IEEE International Conference on Data Mining Workshops (ICDMW)**, pp. 726–733. IEEE, 2024.

[5] Mengyao Cui, et al. Introduction to the k-means clustering algorithm based on the elbow method. **Accounting, Auditing and Finance**, Vol. 1, No. 1, pp. 5–8, 2020.

[6] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Sebti Foufou, and Abdelaziz Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. **IEEE transactions on emerging topics in computing**, Vol. 2, No. 3, pp. 267–279, 2014.

[7] Sara J Weston, Ian Shryock, Ryan Light, and Phillip A Fisher. Selecting the number and labels of topics in topic modeling: A tutorial. **Advances in Methods and Practices in Psychological Science**, Vol. 6, No. 2, p. 25152459231160105, 2023.

[8] Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. Large language models enable few-shot clustering. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 321–333, 2024.

[9] 林和希, 参木裕之. 健康経営度調査テキストに対する定量評価およびレコメンドアルゴリズムの提案. In **Proceedings of the Thirty-first Annual Meeting of the Association for Natural Language Processing**, 2025.

表 3: Wilcoxon 符号付順位検定結果
(Holm 補正後の p 値と効果量を示す。***: p<0.001, *: p<0.05, n.s.: 有意差なし。
手法 1・2 の列では、効果量に基づき優位な手法を太字で示した。)

検定対象	手法 1	手法 2	p 値 (Wilcoxon+Holm 補正)	rank-biserial correlation (効果量)
クラスタ数推定誤差	**提案手法 1&2**	BERTopic	0.012(*)	-0.533
クラスタ数推定誤差	**提案手法 1&2**	従来手法 1&2	**<0.001** (***)	-0.702
ARI	**提案手法 1**	従来手法 1	**<0.001**(***)	0.802
ARI	**提案手法 2**	従来手法 2	**<0.001**(***)	0.800
ARI	**提案手法 1**	BERTopic	0.001(*)	0.602
ARI	**提案手法 2**	BERTopic	0.001 (*)	0.563
NMI	**提案手法 1**	従来手法 1	**<0.001**(***)	0.929
NMI	**提案手法 2**	従来手法 2	**<0.001**(***)	0.966
NMI	**提案手法 1**	BERTopic	0.168(n.s.)	0.305
NMI	**提案手法 2**	BERTopic	0.168(n.s.)	0.315

A　トピック分類問題における各手法の性能に対する検定

表 1 の各データセット対して提案手法および従来手法を適用した際のクラスタ数推定誤差および ARI、NMI に有意差があるのかどうかを以下のフローで検証した。

1. 全体差の検定手法間の差を評価するため、反復測定デザインに適したノンパラメトリック検定である Friedman 検定を実施した。
2. 多重比較の実施 Friedman 検定で有意差が認められた場合、各手法間のペア比較を Wilcoxon 符号付順位検定で行い、Holm 法による多重比較補正を適用した。

この際、クラスタ数推定誤差については**絶対値に変換**した上で検定を行った。

A.0.1　Friedman 検定

クラスタ数推定誤差、ARI、NMI のいずれにおいても、p 値は**<0.001** であり、全体に有意差が認められたため、後続の Wilcoxon 符号付順位検定を実施した。

A.0.2　Wilcoxon 符号付順位検定

検定結果を表 3 に示す。なお、提案手法はすべて「手法 1」の列に記載している。

クラスタ数推定誤差については、提案手法（LLM）は従来手法（BERTopic および Elbow 法）と比較して有意に小さい傾向を示し、効果量も負の値となった。

ARI に関しては、提案手法と従来手法、特に BERTopic との間に有意差が認められた。効果量の観点からも、提案手法（手法 1）は ARI が高くなる傾向がある。

NMI については、提案手法と Elbow 法との間に有意差が認められたが、BERTopic との間には有意差はなかった。ただし、効果量を見ると、提案手法は NMI においても相対的に高い値を示す傾向がある。

B　実験に使用したシステムプロンプトおよびレスポンススキーマ

```
あなたはデータ分析の専門家です。
以下のテキストデータセットに対して、
クラスタリングを行う際に適切と思われる
クラスタ数 (recommended_number_of_clusters: int)
を出力してください。

# 入力情報
- 類似度が高いものが隣接するようにソート済み
- データ件数: N

# 注意点
- クラスタ数は意味的に一貫したグループを形成できるように選んでください。
- クラスタ数が多すぎると過分割、少なすぎると異質なデータが混在します。
- データ件数と多様性を考慮してください。
```

図 3: システムプロンプト

Listing 1: LLM 応答形式のレスポンススキーマ

```
{
  "title": "ans_format",
  "type": "object",
  "properties": {
    "recommended_number_of_clusters": {
      "title": "Recommended Number Of Clusters",
      "type": "integer"
    }
  },
  "required": ["recommended_number_of_clusters"]
}
```