情報量最大化クラスタリングによる次元削減を用いた文書の類似度計算

Document Similarity Calculation Using Dimensionality Reduction by Invariant Information Clustering

2024Y14　相撲　美月（Sumai Mitsuki）
担当教員　出口　利憲（Deguchi Toshinori）・堀内　咲江（Horiuchi Sakie）

1. 序論

急速な情報化により膨大なデータが活用される一方、SNS の普及で信頼性の低い情報も拡散し、情報リテラシーが重要視されている。そのため、統計学や人工知能を活用したデータマイニングが注目されており、特に自然言語処理を活用したテキストマイニングは、文章から法則性や相互関係を見出す技術として活用されている。しかし言語の意味を正確に捉える難しさからさらなる研究が必要とされている。

このテキストマイニングにおける文書分類という分野では、文書間の類似度計算を効率化するために次元削除が用いられる。特徴ベクトルの次元を減らすことで、計算量を抑えつつ、データの意味を保ちながら新たな特徴量を抽出できる。従来の次元削減手法として、潜在的意味解析（LSA）等の統計的手法が挙げられる。単語の出現回数などに基づき次元を削減することで、文書の関係を把握する手法であるが、単語の意味が十分に考慮されないという課題がある。また従来のクラスタ分析は、類似するデータ同士をグループ分けするものであるが、分析対象の次元数やデータ量の増加に伴い、計算負荷が増大し、クラスタリングが困難になることが指摘されている。これらの課題を解決するため、情報量最大化クラスタリング（IIC）[1] が新たなクラスタ分析手法として提案された。本研究では、この手法を単語のクラスタリングに応用することで、文書間の類似度を高精度に計算できると予想している。

2. 目的

本研究の目的は、次元削減という技術において、新たな手法として IIC の利用を提案・実装し、その有用性について検討することである。本手法は従来の教師なし手法でのクラスタが１つにまとまる又は消失する現象と、クラスタの特徴量が他のクラスタと類似する現象を解決できる手法として注目されている。この手法と、従来の次元削減手法との比較を行い、提案手法の有効性を検証する。

本研究では、Word2vec によって得た単語のベクトル表現に対して IIC を行い、各単語をそのクラスタへ属する確率値が最大のクラスタに分類する。各文書に含まれる単語が、どのクラスタにどれほど属しているかを集計する。これによって得られた単語とクラスタの関係を用いて、文書行列を次元削減し、各文書間の類似度計算を行う。また LSA などの次元削減手法での類似度計算の結果と比較する。評価方法としては、文書間の類似度に基づき文書を分類し、分類した文書が正しいジャンルに一致しているかを検証する。

3. 実験方法

実験は以下の手順で進めた。

- データの取得と前処理
- 単語の TF-IDF 値の計算、IIC
- 次元削減
- 結果の比較

本科在学時の卒業研究では、複数の文章から抽出した単語集合に対して IIC を実装し、この手法の有用性、優位性を確認した。各クラスタ内の単語を確認した結果、ニューラルネットワーク内のパラメータ調整が精度に大きな影響を与えることが示されたものの、階層的クラスタリング（ウォード法）での分類と比べて、より類似度の高い単語を同一のクラスタに分類できることを確認した。

特別研究 1 では、文書分類のためにニューラルネットワークのパラメータを学習した IIC を用いて次元削減を行い、文書間の類似度計算を行った。またその評価のために文書の分類を行った。

IIC でのニューラルネットワークは全結合層二層で構成した。一層目のユニット数とバッチ正規化を行う層のユニット数は実験で変化させ、活性化関数に ReLU 関数を用いている。二層目のユ

ニット数はクラスタ数となり、softmax 関数を用いてデータがどのクラスに属するかの判断を行う。また、オーバークラスタリング用の二層目のユニット数は、オーバークラスタリング率にクラスタ数を乗算したものとした。IIC における入力には、元のベクトルデータと元のベクトルデータに全ベクトルデータの標準偏差に基づくノイズを加えたものを使用する。ネットワークにこれらを入力した際の、それぞれの出力の相互情報量が最大となるように学習を行う。

実験では、各文章に含まれる特定の単語がその文書全体でどのくらい重要かを表す TF-IDF 値を用いて、文書数×単語数の文書行列を作成する。また、Word2vec によって得た単語のベクトル表現に対して IIC を行い、各単語を各クラスタへ属する確率値（推定確率）が最大のクラスタに分類する。この値を各文書が含む単語がどのクラスタにどれほど属しているかを表す帰属度として使用し、分類したクラスタ以外の帰属度は 0 とする。これにより単語数×クラスタ数の行列を作成する。これらの行列を掛け合わせることで、次元削減した文書行列が求められ、文書数×クラスタ数の行列を生成する。文書間の類似度は、次元削減された行列に対して、２つのベクトルがどのくらい似ているかという類似性を表す尺度の１つである cos 類似度を用いて計算する。類似度の計算結果は、図に示すデンドログラムにて可視化する。

類似度計算の評価のための文書分類として、文書間の類似度を基にクラスタ分析を行った。距離の測定方法にウォード法を指定した階層的クラスタ分析を実行した。出力した分析結果に対して fcluster 関数を使用し、実験パターンごとに割り当てられたジャンル数分までクラスタの分割を行った。ジャンル数分に分割されたクラスタに対して、ラベルの割り当てを総当たりで調べ、最大の正解率を算出した。

4. 実験結果

Fig. 1 は、livedoor ニュースに現在掲載されている記事のうち、「スポーツ」「食べ物」「産業」「教育」「気象」の 5 つのジャンルから 2 個ずつ合計 10 個の文書を収集した際の類似度の計算結果である。IIC はオーバークラスタリング率が 100、ユニット数が 400 のものを用いて実験を行った。Fig. 1 より、食べ物とスポーツと教育の文書は類似度が高いことが分かった。気象の文書の類似度が低い原因は、気象の記事では気温や日付といった数値が多いが、数値がベクトル表現できないことで、文書中の必要な情報が減ったためだと考えられる。そのほかにも、ジャンル数と文書数を変えて実験を行った。それぞれ正解率を算出し、その一部を抜粋したものを Table 1 に示す。Table 1 より、おおよそ 60〜75 %の正解率であることが分かった。

5. まとめ

現段階では、IIC による次元削減を用いた場合と LSA 等の次元削減手法での文書の類似度計算の比較が出来ていないため、手法の有効性を確認できていない。また本科在学時の卒業研究にて、全結合層の各層の出力ユニット数の値が IIC の結果に影響を与えることが分かったが、今回の文書分類での各パラメータの適切な値について十分な検証ができていない。さらに実験に用いたデータ数が少ないため、文書数を増やして検証を行う必要がある。特別研究２においては、この点について実験を行いたいと考えている。

参考文献

[1] Xu Ji, Joao F. Henriques, Andrea Vedaldi,「Invariant Information Clustering for Unsupervised Image Classification and Segmentation」, IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, p.9865-9874

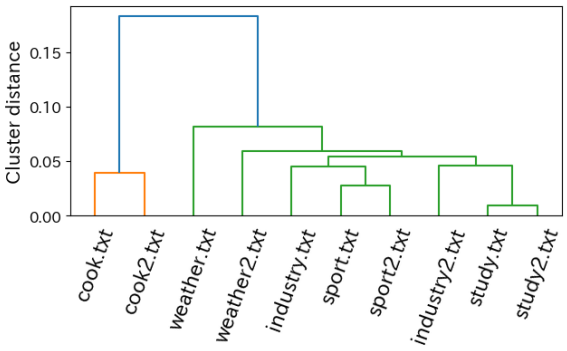

Fig. 1 Similarity calculation results.

Table 1 Accuracy

ジャンル数	文書数	正解率 [%]
5	10	70.0
5	15	73.3
6	12	66.7