研究報告

超高速カテゴリカル特徴選択手法を用いた高次元データの特徴付け

学習院大学　計算機センター　久保山　哲　二
兵庫県立大学大学院　応用情報科学研究科　申　吉　　浩
千葉商科大学　商学部　橋　本　隆　子
学習院大学　計算機センター　城　所　弘　泰
学習院大学　計算機センター　磯　上　貞　雄
学習院大学　計算機センター　村　上　登志男

1　はじめに

コンピューターネットワークのログ、Web ページ上のテキストやリンク、遺伝情報をはじめとして、多種多様なデータが日々蓄積されている。しかし、これらの膨大なデータは、一般に多次元データである。たとえば、人のデータには、(氏名, 性別, 生年月日, 身長, 体重, …) 等の様々な属性があり、属性数の次元をもったデータと捉えることができる。また、テキストデータには、1 つの文書あたり、多数の単語が出現するため、単語数の次元をもったデータと捉えることができる。

このような多属性・多次元のデータの中から目的に応じて必要な属性・次元を取り出すプロセスを「特徴選択」といい、データ分析において非常に重要な役割を果たす。例えば、大量の遺伝情報のなかから、特定の疾患に関係のある配列の組合せを取り出すことができれば、疾患の遺伝的因子の特定に繋がる。また、膨大な文書の中から、ある文書を特徴づける単語を取り出すことができれば、大量文書の要約やトピック抽出に応用することができる。

本研究では、申請者が研究メンバーの申らと開発した超高速なカテゴリカル特徴選択アルゴリズム処理系 [1] のこれまでにない新しい応用領域を探ることが目的である。

2　特徴選択手法の応用

本研究の一環として、東日本大震災直後のツイッターデータからトピックを抽出する実験を行った。対象データは震災発生時刻周辺の約 2 億件のツイートである。このデータにおけるツイッター ID 数は約 100 万である。現在トピック抽出においては、LDA(Latent Dirichlet Allocation) が広く用いられているが、今回の規模のデータに対しては計算量が膨大になり適用困難である。

この問題に対応するため、本研究では、高速な特徴選択アルゴリズム CWC を適用する。CWC は教師付きのアルゴリズムであるため、クラスラベルの付与が必要となる。そこで、まず k-means 法によるクラスタリングを施した後、各クラスタをクラスラベルとして、各々のクラスタを特徴づ

ける少数の単語を抽出した。

CWC の処理系については、スパースデータに対応した高速な処理系 sCWC を以下のサイトにて公開している。

> Tetsuji Kuboyama, sCWC: very fast feature selection for nominal data, `https://github.com/tkub/scwc`, (2017 年 8 月現在)

また、この応用の詳細については、以下の国際会議ワークショップで採択され報告済みである。

> Takako Hashimoto, Dave Shepard, Tetsuji Kuboyama, Kilho Shin: Topic Extraction Method from Millions of Tweets Based on Fast Feature Selection Technique CWC, 2016 IEEE International Conference on Data Mining Workshops, pp.724–731, 2016.

3 おわりに

大量データからのトピック抽出に、特徴選択アルゴリズム CWC の有用性を示すことができた。しかし、今後、適用領域拡大のために、さらに以下のような課題にとりくむ必要がある。(1) 特徴選択により選択された特徴と強い相関を持つ特徴は選択されないため、トピック抽出において、類似性が高くトピックの解釈に寄与する単語が選ばれない可能性が高い。そのため、トピックのコヒーレンス尺度の観点からよい単語集合を抽出する仕組みを組み込む。(2) クラスラベルを説明するための特徴の選び方は、1 通りではないが、CWC は深さ優先探索により最初に見つかった極小特徴集合を解として出力するため、複数のトピックが抽出できる場合でも、そのトピックを取りこぼしてしまう可能性がある。そこで、CWC に他の極小解を探索する仕組みを組み込む。

参考文献

[1] Kilho Shin, Tetsuji Kuboyama, Takako Hashimoto and Dave Shepard: Super-CWC and super-LCC: Super fast feature selection algorithms, 2015 IEEE International Conference on Big Data, pp.61–67, 2015.