平成27年度　研究成果発表会

確率的モデリングによるビッグデータ活用

○櫻井 瑛一 *1) 本村 陽一 *1)

■キーワード　ビッグデータ活用、確率的機械学習

1. PLSA によるデータのクラスタリング技術紹介
2. Bayesian Network によるデータ解析技術紹介
3. 両者を組み合わせることによるクラスタ意味解析

■概要

インターネットの発展とともにデータを収集し、データを生成している人間の行動を理解しようという試みがなされてきた。こうして始まったビッグデータ活用の流れは、データを中心とした分析の重要性を明らかにし、分析の背景技術としての機械学習手法研究の推進を後押ししてきた。この機械学習の中でも確率を基礎とする手法は、人の行動の多様さもあって、実際のデータによく適合するため、文書分類をはじめとした多くの分野への適用と手法の大きな発展を遂げてきた。

本発表では、この確率的手法の中でも潜在的確率意味解析（以下、PLSA とする）と Bayesian Network（以下、BN とする）の紹介とこれらをインターネット上のデータではない実際のデータへの適用により、インターネットのビッグデータだけでなく、現実社会のデータに対しても確率的モデリングは非常に有用であることを示す。

■技術概説

(1) 潜在的確率意味解析 (PLSA)[1]

PLSA とは、大量の文章の分類を行う一手法である。この手法では、文書内の単語が出現する確率がその話のテーマによって変化する、という確率的モデルを考え分類を行う。図 1 がそのモデルの概略図である。

この手法自体は、文書分類のみならず、アンケートデータや購買データなどの文書以外のデータでも適用可能であり、回答や購買行動の典型的傾向が似た人を分類することが可能となる。

(2) Bayesian Network(BN)[2]

ある条件の下 (X) でのある現象 (Y) が生じる確率を条件付確率と呼ぶ。この、条件付確率では Y という状況に X が及ぼす、と解釈ができるため、X → Y と矢印を張ることにより、因果構造を示すことができる。

この矢印のネットワークによって表現されるグラフが BN である（図 2）。

(3) 両者の組み合わせによる、データのクラスタリングと意味解析[3]

PLSA による分類は、人の潜在的傾向をあらわにすることが可能であるが、各々の分類が何を特徴としているかは実データの場合分かりにくい場合が多い。そこで、因果関係を表す BN と PLSA による分類結果を組み合わせることで、PLSA による分類の説明モデルが作成できることを本発表では示す。そして、その分類モデルがどのような企業の課題解決に役立ったかを示す。

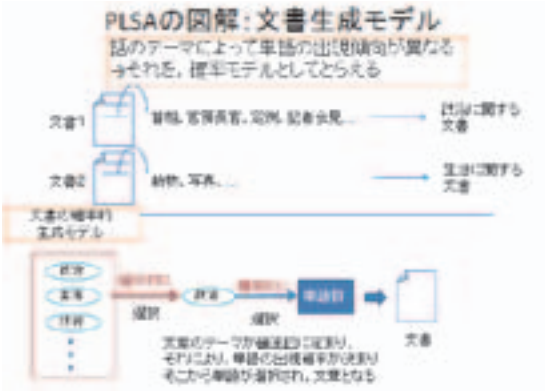

図 1. PLSA 概説図

図 2. BN の例

参考文献
[1] T. Hoffman, Machine Learning, Vol.42, pp.289-296 (1999)
[2] J. Pearl, In Proceedings of Cognitive Science Society, pp.329-334 (1985)
[3] 石垣等，人工知能学会全国大会，3J1-NFC1a-2 (2010)

*1) 国立研究開発法人産業技術総合研究所