自然言語における分類モデルの学習に関する研究

加瀬, 雄一朗 / Kase, Yuichiro

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
58
(開始ページ / Start Page)
1
(終了ページ / End Page)
2
(発行年 / Year)
2017-03-31
(URL)
https://doi.org/10.15002/00014245

自然言語における分類モデルの学習に関する研究

STUDIES ON LEARNING CLASSIFICATION MODELS IN NATURAL LANGUAGE

加瀬　雄一朗
Yuichiro KASE

指導教員 三浦孝夫

法政大学大学院理工学研究科システム理工学専攻修士課程

Along with the increase in the amount of data, the importance of automatically processing them efficiently on computers is increasing. In this Investigation, we deal with classification which is one of essential technologies for automatic processing by computers. Specifically, by constructing classification models that consider the characteristics and properties of the data, it is possible to accurately and speedily attempt to perform automatic classification. Our Experiments also show the effectiveness of each classification model.
Key Words *: Natural Language Processing, Classification, Data Mining*

1.　問題の背景

近年，スマートフォン等の情報機器やインターネットの普及を背景として，様々な分野のタスクが電子化されており，それに伴い電子的に表現されたデータの量が増大している．例えば，新聞社の多くは紙版のみならず電子版の新聞も扱っており，現在，ウェブ上には大量の新聞記事のデータが存在する．電子化された情報を保持することは，紙媒体の情報などと比べると，維持・管理や検索などが比較的容易であり，実際，世の中の電子的なデータの量は増え続けている．そのような大量のデータを，機械で効率的に自動処理することの重要性はますます高まっている．例えば，多くの企業は，このような大量のデータを “ビッグデータ” と称し，それらを利用して，顧客の満足度を向上させることを試みている．

大量のデータを機械で自動処理する際，データの自動分類は必須技術のひとつである．データを自動的に分類することで，目的のデータへ簡単にアクセスすることが可能になる．例えば，日々大量に生成される新聞記事を，その記事のジャンルごとに自動分類できれば，後に “経済” に関する記事のみを参照したいといった場合などに，必要な記事をすばやく得ることができる．

自動分類を行うには，想定するデータ領域に対応したモデルを構築する．パラメタをもつモデルを構築し，そのパラメタを具体的なデータに合わせて調整 (学習) することで，そのデータ領域に対応した特徴を汎用的に表現することができる．そして，各カテゴリに対応したモデル (分類モデル) をあらかじめ構築・学習しておき，分類対象事例の特徴が，どのモデルの特徴に一番近いかを判断することで分類を行う．

分類モデルを構築するにはいくつかの問題がある．まず，分類性能が学習データの質および量に依存することである．しかし，学習データを作成するコストは大きい．なぜなら，分類は人間が事前に決めた基準にしたがって行われるため，人手でその基準に従ってラベル付けされたデータ (学習データ) を作成する必要があるからである．また，別の問題として，各事例が複数のカテゴリに同時に所属することがある．例えば，「オリンピック招致に関する政府の取り組み」に関する新聞記事があるとする．この記事の分類を考えた場合，少なくとも，“政治”，“経済”，“国際” のジャンルに同時に分類されるべきであり，“政治”，“経済”，“国際” のそれぞれのクラスに排他的に分類することはできない．さらに，各分野に依存した情報には，テキスト，音声，画像など多様なデータ表現があり，データ表現に応じて異なる性質を有することがある．この場合，それぞれのデータの性質に応じたモデル構築が必要となる．

本研究では，データの自動分類を効率的かつ高速に行う分類モデルを提案する．まず，データ中に潜在的に存在するクラスの発見手法を提案する．加えて，ラベル間の依存性を確率的に考慮し，ひとつのカテゴリには分類できないデータ，特に，文書データに対して，同時に複数のカテゴリに分類する新たな手法を提案する．さらに，系列データのモデル推定を高速に行う手法を提案する．また，これらの分類モデルを実際のデータに対して適応することで，提案手法の有効性を確認し，精度および実行効率のいずれも優れた結果を得ることができることを示す．

2.　扱う問題

(1)　線形混合モデルによるニュースコーパスの多重ラベル分類

本研究では，ニュースコーパスに内在する潜在クラスの抽出手法を論じる [1]．主たるアイデアは，基本となるラベルベクトルと潜在クラスとの関連を見出し，ベクトル集合を潜在クラスに対応させることにある．各カテゴリの概念は，複数のラベルが表す概念の重み付けの組み合わせで表現できる場合があるが，どのカテゴリも独立した意味を表している．即ち，複数のラベルを組み合わせて潜在クラスを定義できると考える．

具体的には，まず，所与のラベルに従ってデータを多重ラベル分類し，多項分布上で準学習方式によりラベル所属確

率を求め，ラベルベクトルを生成する．次にラベル空間上でベクトルをクラスタリングし，潜在クラスと対応付ける．確率値を推定するため，ここでは学習データおよび未分類文書データの両方を使用する．この確率ベクトルからなるクラスタを，潜在的なクラスと考え，文書に対して当該クラスに分類するとみなす．

(2) 多重同時関係を考慮した多重ラベル分類

本研究では，文書の自動分類手法，特に各データが複数カテゴリに同時に所属する場合における分類手法を扱う．本稿の主たるアイデアは，ラベル間の依存性を確率的に考慮することにある [2][3].

具体的には，同時に付与されやすいラベルの組合わせを，データ集合に頻出するラベルの組合わせであると仮定し，学習データ中に頻出するラベル間の同時関係を，データマイニングにより，直接抽出する．その際に，学習データ中に出現する，データの構成要素間の関係，ラベル間の関係，およびデータ集合とラベル集合の間の関係の 3 つの同時関係を，一度に抽出する．そして，これらの多重同時関係を用いて，確率密度関数を推定することで，確率的にラベル間関係を考慮した多重ラベル分類を行う．

(3) 制限付き隠れマルコフモデルによる系列モデルの高速推定

本研究では，系列データのモデル推定を高速に行う手法を提案する [4]．系列データとは，データを構成する要素が連続し，その要素の出現が互いに関連しあうような性質をもつデータである．系列データをモデル化するため隠れマルコフモデル (Hidden Markov Model, HMM) の状態構造 (トポロジ) に対して，系列中で「状態遷移では後戻りしない」という制限を加えることで，モデルの学習の効率化を行う．

具体的には，ベイジアンネットワークにも用いられるグラフ構造である有向非巡回グラフ (Directed Acyclic Graph, 以下 DAG) に自己サイクル (同じ状態での繰り返し) を追加した構造を HMM の潜在状態構造に用いて，系列データのモデル化を試みる．また，Left to Right HMM と同様に初期状態と最終状態を一つずつ持つと仮定する．以上の潜在状態の構造に対する制限により，前向き確率，後ろ向き確率の計算時に必要な計算量を大幅に削減することで，学習時計算の効率化を実現する．

3. 結果

はじめに，線形混合モデルによるニュースコーパスの多重ラベル分類では Reuter-21578 を用いて実験を行う．その結果，多重ラベルが完全一致した場合の正解率は 0.669 である．また，再現率は 0.745，適合率は 0.770 である．単純ベイズ法との比較では，再現率はほぼ全てのラベルで単純ベイズが優れているのに対し，適合率は逆に提案手法が全てのクラスで勝り，全体としても 748% の劇的な改善がみられる．

多重同時関係を考慮した多重ラベル分類でも同様に Reuter-21578 を用いて実験を行う．その結果，多重ラベルが完全一致した正解率は 0.840 であり，単純ベイズ法との比較で，平均 0.065(108%) の改善である．また，再現率は平均で 0.813，適合率は平均で 0.740 であり，平均再現率は単純ベイズのほうが 0.032(104%) 値が優れているが，平均適合率では提案手法のほうが 0.036(105%) 値が優れている．

そして，制限付き隠れマルコフモデルによる系列モデルの高速推定については，そのモデル化の妥当性を示すため，提案手法を分類問題に適用する．The 20 Newsgroups Data Set を用いて実験を行った結果，提案手法の学習の実行時間は平均で 73.917ms であり，Ergodic-HMM との比較で平均 29.6%である．また，平均正解率は 0.7631 であり，Ergodic-HMM との比較で，平均 0.0118 の改善である．図 1 に制限付き隠れマルコフモデルより抽出された構造例を示す．

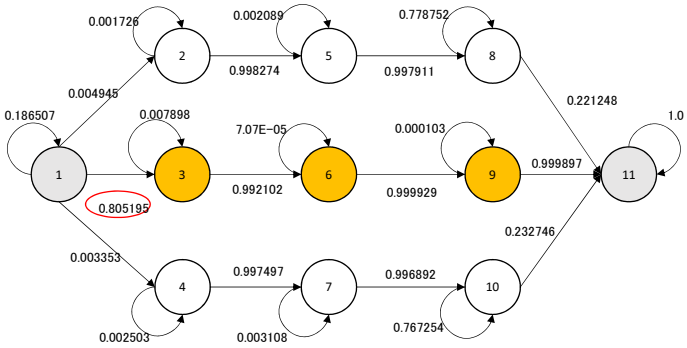

図 1 抽出された構造例

4. 結論

本研究では，データを計算機で自動分類することを目的として，主に多重ラベル分類と系列データのモデル化方法を提案し，それぞれの分類モデルが特に文書データに対して有効な手法であることを実験により示した．

線形混合モデルによるニュースコーパスの多重ラベル分類では，ニュースコーパスに内在する潜在クラスを抽出し，少量のラベル無しデータと大量のラベル付きデータを用いて，ラベル情報を更新しながら，有効な多重ラベル分類が行えることを示した．多重同時関係を考慮した多重ラベル分類では，ラベル間の依存性を確率的に考慮することで，学習データ中に出現するデータの構成要素間の関係，ラベル間の関係，およびデータ集合とラベル集合の間の関係を扱い，高速かつ効率的な多重ラベル分類が行えることを示した．制限付き隠れマルコフモデルによる系列モデルの高速推定では，系列データをモデル化するため隠れマルコフモデルの状態構造に対して，系列中で「状態遷移では後戻りしない」という制限を加えることで，モデルの学習の効率化を行えることを示した．

参考文献

1) Yuichiro Kase, Takao Miura. “Mining Classes by Multi-label Classification.” EGC. 2015.
2) Yuichiro Kase, Takao Miura. “Multi-label classification using labelled association.” Communications, Computers and Signal Processing (PACRIM), 2015 IEEE Pacific Rim Conference on. IEEE, 2015.
3) Yuichiro Kase, Takao Miura. “Probabilistic Classification Using Data Mining.” Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2015.
4) 加瀬雄一朗, 三浦孝夫. “制限付き隠れマルコフモデルによる系列モデルの高速推定” 第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM 2017), 2017.