論文内容の要旨

博士論文題目　Text Categorization Using Machine Learning
(機械学習を用いたテキスト分類)

氏名　　平 博順

(論文内容の要旨)

近年、大量のテキストデータが利用可能になったことや、コンピュータの性能が大幅に向上したことから、機械学習的アプローチ、すなわち人手によりカテゴリラベルを付与したテキストデータから自動的に分類器を作成する方法が、分類精度・省力性・保守性に優れているために主流となっている。本論文では、機械学習を用いたテキスト分類に関して、次のような点について研究を行なった。すなわち、大量の単語属性を用いて高精度の分類を行う方法、分類ラベル有り訓練データが少ないときにラベルなしデータの分布を利用する方法、および、下位カテゴリの存在を仮定し、能動的ラベル付けを用いて効率よく高精度の分類を実現する方法である。

機械学習を用いて高精度のテキスト分類を行うためには、一般に沢山の単語属性が必要であり、従来の学習手法では沢山の単語属性を入力として使用すると訓練データに対して必要以上に適合して、未知のデータに対する分類精度は落ちるという過学習が起きる問題があった。そこで様々な情報量で属性選択を行い、学習に使用する属性の次元を数百程度まで減らして学習を行っていた。しかし、数百程度の属性では、十分に高い分類精度を持つ分類器を学習することは困難であった。そこで、過学習しにくいとされる新しい機械学習手法であるサポートベクタマシンを用いてテキスト分類をおこなった。さらに、サポートベクタマシンを用いる場合にも属性選択は有効であるかどうかを評価した。

次に、トランスダクティブ・ブースティングを用いたテキスト分類について研究を行った。訓練データが十分に得られる場合には、サポートベクタマシンやブースティングなどのマージン最大化分類は高い分類精度を持つ分類器を生成するために有効な手法であるが、訓練データの作成は人手で行うため、コストが高く、学習に十分な訓練データが得られないことも多い。そこで、分類ラベルが付与されていない未知のデータの分布も考慮に入れて学習を行うトランスダクションの手法をテキスト分類の学習に取り入れた。サポートベクタマシンと同じマージン最大化分類器の一つであるブースティング、特にアダブーストアルゴリズムにトランスダクションを組み入れる方法を提案した。

さらに、拡張結合混合分布モデルを用いたテキスト分類についての研究を行った。与えられたカテゴリの中に潜在的に下位カテゴリを仮定したときの分類について述べる。さらに、訓練データが少ないときに、未知データに対して、分類に重要だと思われる少数のデータを機械的に選択、提示し、少数のデータに人手でカテゴリ付与を行うことで、効率よく高精度の分類器を得る「能動的ラベル付け」手法を提案した。

氏 名	平 博順

（論文審査結果の要旨）

平成13年12月25日に開催した公聴会の結果を参考に平成14年2月12日に本博士論文の審査を行った。以下のとおり、本博士論文は、提案者が独立した研究者として、研究活動を続けていくための十分な素養を備えていることを示すものと認める。

平 博順は、本博士論文において、機械学習を用いた文書分類についていくつかの新しい方法を提案した。特に、次のような視点に基づいて、研究を行い、高い分類精度が達成できることを示した。

1. サポートベクターマシンを用いた文書分類において、どのような基準によって学習に用いる属性（単語）の選択を行うべきかについて考察し、いくつかの方法を提案した。提案手法を実データによって検証し、頻度による選択や情報量に基づく選択等を比較検討し、それらよりも品詞に基づく単語選択の方がよい精度を達成することを示した。

2. 訓練データが十分に得られる場合には、サポートベクタマシンやブースティングなどのマージン最大化分類は高い分類精度を持つ分類器を生成するために有効な手法であるが、訓練データの作成は人手で行うため、その作成に要するコストが高い。そこで、分類ラベルが付与されていない未知のデータの分布も考慮に入れて学習を行うトランスダクションの手法をテキスト分類の学習に取り入れた。サポートベクタマシンと同じマージン最大化分類器の一つであるブースティング法の一つであるアダブーストアルゴリズムにトランスダクションを組み入れる方法を提案し、その効果を実証した。

3. 拡張結合混合分布モデルをテキスト分類に適用することを提案し、その有効性について検証した。この問題を、与えられたカテゴリの中に潜在的に下位カテゴリを仮定したときの分類として定式化した。また、訓練データが少ないときに、未知データに対して、分類に重要だと思われる少数のデータを機械的に選択し、少数のデータに人手でカテゴリ付与を行うことで、効率よく高精度の分類器を得る能動的ラベル付け手法を提案し、その効果についても検証した。

文書分類に対する機械学習方式を様々な視点から検討し、いくつかの新しい手法を提案した本研究は、独創性が高く、しかも実用的であり、自然言語処理の分野において高い貢献があると評価する。

よって、本論文は博士(工学)の学位論文として価値あるものと認める。