論文内容の要旨

博士論文題目　Clustering Approaches to Text Categorization
(文書分類へのクラスタリングによるアプローチ)

氏名　　高村 大也

(論文内容の要旨)

本論文は，文書分類の精度を向上させる方法を提案したものである．文書分類とは，文書をその内容に従って分類するタスクであり，電子メール分類やウェブページ分類など実用上重要なタスクの一般的な枠組と捉えられる．この目的に対する様々なアプローチのうち，本論文ではクラスタリングによるアプローチを論じている．クラスタリングは一般的には，教師無し学習の方法であるが，文書分類の精度を向上させるために利用できることを示している．

第一のアプローチは，単語と文書の共クラスタリングによる方法である．分類問題に対する確率モデルのアプローチでは，一つのカテゴリに一つの確率分布を仮定する場合が少なくない．このアプローチの動機づけとして，この仮定が正しいとはいえないことを実験的に示した．続いて，この仮定による確率モデルの乱れを是正する枠組を，文書と単語の共クラスタリングに基づいた形で提案した．提案方法では，確率モデルの是正のために文書がマージされるが，データスパースネス問題を軽減するために単語のクラスタリングも同時に行う．この方法を用いて文書分類の精度を向上させることに成功した．

第二のアプローチは，クラスタリングに基づいた構成的帰納学習法である．本手法においては，サポートベクターマシンと呼ばれる分類器が，潜在的意味解析 (LSI) などのクラスタリングに基づいた構成的帰納学習法と組み合わせて使用される．クラスタリングによって抽出された素性が文書の素性空間の拡張を実現しているこの手法は，テストデータ中の素性で訓練データに出現しないものが多く存在するという文書データの性質を利用したものである．クラスタリングに使用できる未知データが充分多く存在する場合に分類精度が向上することが示された．

最後に，確率分布に基づいたカーネル関数を文書分類に応用する試みについて述べている．ここでは，文書データが複数のカテゴリの混合であることを利用して確率分布を作る．二値分類の場合においても，このような確率分布から構成したカーネル関数が有効であることが示された．

氏 名	高村 大也

（論文審査結果の要旨）

平成15年1月28日に開催した公聴会の結果を参考に平成15年2月17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．

高村 大也は，本博士論文において，クラシタリングに基づき以下に示すように様々な手法を利用することによって文書分類の性能向上をを行う方法を提案した．

1. 単語集合のクラスタリングと文書集合のクラスタリングを同時に行なう共クラスタリングという手法を提案し，文書分類に有効に働くことを示した．また，これにより，従来単一の確率分布よりなると仮定されることが多かった文書カテゴリに対し，カテゴリによっては複数の部分カテゴリを考えることが有効であることを示した．

2. 情報検索で用いられる潜在的意味解析のような次元圧縮法は，文書データを表現する単語空間に対するクラスタリングの一種と考えられる．これによる軸を追加素性として加えることにより学習性能が向上することを示した．

3. 確率分布に基づいたカーネル関数の一つであるTOPカーネルを文書分類に応用する試みを示した．文書データが複数のカテゴリの混合であることを利用し，確率分布を作り，二値分類の場合においても，このような確率分布から構成したカーネル関数が有効であることを示した．

様々なクラスタリングに基づく素性選択によって文書分類性能の向上のための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，文書処理の分野において高い貢献があると評価する．よって，博士(工学)の学位論文として価値あるものと認める．