審 査 の 結 果 の 要 旨

氏　名　嶋　幸太郎

本論文は、テキスト分類などの高次元データの分類問題に対し、サポートベクターマシンを用いた属性選択法に着目し、同手法の高次元データへの有効性について論じたものである。本論文は、全７章で構成されている。

第１章では、研究の背景が述べられている。近年のデータ量の増加とともにデータを効率的に分類する技術が重要となってきているが、計算コスト・分類精度・モデルの解釈のし易さといった観点から、属性選択・抽出などの次元圧縮を行い、分類に寄与する属性（有効分類属性）を特定することが望ましい。従来の次元圧縮手法は比較的少数の属性数を想定していたが、それらの手法は現在扱われるような数万以上にも上る属性数には十分に対処しきれない。そこで、高次元データに対する効率的な次元圧縮手法が必要であるという本研究の動機付けを行っている。

第２章では、テキスト分類について概説している。テキスト分類で用いられる、ベクトル空間モデル、前処理過程、アルゴリズムの分類精度の評価指標について説明している。そして、テキスト分類で従来用いられてきた次元圧縮手法についてまとめ、それらの問題点を指摘している。

第３章では、本手法で着目する、サポートベクターマシンの原理、およびサポートベクターマシンを用いた属性選択法についての説明を行っている。同手法の利点、既往の研究で未だ十分な知見が得られていない課題点についてまとめている。

第４章では、同手法を標準的なテキスト分類データセットに適用している。属性選択に関して従来手法との比較を行い、同手法が多変量性のために、従来手法よりもクラスに依らず頑健に不要属性を除去できることを示し、不要属性数を推定する指標を提案している。さらに、ソフトマージンパラメータが同手法に及ぼす影響についての分析を行っている。ソフトマージンパラメータの設定値によって属性選択の効果が大きく異なることを明らかにし、同パラメータが大きい場合には不要属性の特定に適しており、小さい場合には少数の有効分類属性の特定に適していることを示している。その上で、両方のケースの利点を併せ持つ属性選択指標を提案している。そして、本手法を Latent Semantic Indexing (LSI)と呼ばれる属性抽出手法に適用することで、通常の LSI よりも少数の属性数で分類精度の向上を実現できることを示している。

第５章では、本手法をウェブページの情報フィルタリングに適用した結果を報告している。ユーザがウェブページを閲覧し各自の興味の有る／無しを評価した値を元に、本手法を用いてユーザのウェブページに対する興味を表す因子の特定を試みている。その結果、特定された因子とユーザ自身による自分の興味の認識との間に大きな相違が無いことが示され、ユーザの潜在的な興味を表す因子が特定できる可能性が示唆されている。

第６章では、同手法を非線形属性に拡張を行っている。手書き数字認識の分類問題に適用した結果、同手法を用いて非線形な有効分類属性を特定することにより、比較的少数で高い分類精度を実現できることを示している。したがって、同手法が線形属性のみならず非線形属性に対しても有効であることを示している。

第７章は結論であり、本研究で得られた知見がまとめられている。

以上のように、本論文は、サポートベクターマシンと呼ばれる多変量手法が、高次元データからの有効分類属性の特定に有効であることを示した研究であり、データマイニング技術のみならず、データ工学全般への発展に寄与するところが少なくない。よって、本論文は博士（工学）の学位請求論文として合格と認められる。