論文内容の要旨

博士論文題目　Statistical and Graph-based Approaches to Small Sample and High Dimensional Data
(少数サンプルと高次元データへの確率とグラフに基づくアプローチ)

氏名　鈴木郁美

(論文内容の要旨)

近年，機械学習は，さまざまな種類のデータの解析に使われている．機械学習の目的は，新しいデータに対する予測を行うために，既存データを用いてモデルを構築することである．精度の高いモデルを構築（あるいは選択）するためには，訓練データを以下の二つの観点から調べ，対策を立てる必要がある．

(1) 訓練データは，どんなテストデータの予測にも十分対応できるだけの数が集められているか？

(2) 訓練データがベクトル値で表現されている場合，その次元数は「次元の呪い」の影響を受ける程に大きくないか？

(1) に関して本論文で取り上げる問題は，訓練サンプル数の小さい場合の予測モデルの選択である．実際例として，正確な予測が求められる，遺伝子発現プロファイル（マイクロアレイデータ）による癌の診断に注目した．癌など病気の診断に用いる遺伝子発現量プロファイルは，得られるサンプル数が限られるために，実用化には判別器の信頼性が問題になる．交差検定によって複数の判別器の性能を評価し，その評価が最大となる判別器を選ぶことが行われてきた．本論文では，テスト性能の分散を考慮することで，悪い判別器が得られるリスクを回避してモデル選択を行う min-max 法を提案する．提案手法がデータ出現の偏りに起因した性能評価のばらつきに影響を受けにくく，リスク回避型のモデル選択基準として有効であることを示す．

(2) に関して，データの高次元性に関わる問題を扱う．最近，新しい「次元の呪い」として，「ハブの出現」が報告された．本研究では，通勤時間カーネルを始めとするラプラシアンベースのカーネルを類似度がハブを抑制できる可能性がある性質を持つことを示す．実験から，ラプラシアンカーネルは常にではないがハブを抑えることがわかった．ハブが減少したデータセットでは，分類精度も向上した．このことから，教師なしで測れるハブの出現度合いを調べることにより，与えられたデータセットに対するラプラシアンカーネルの有効性を予め評価できると考えられる．

氏　名	鈴木郁美

（論文審査結果の要旨）

平成23年12月19日に開催した公聴会の結果を参考に平成24年2月13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．

鈴木郁美は，本博士論文において，機械学習における2つの問題を取り扱った．一つは，充分な数の訓練サンプルを得ることが困難な場合に，いかに信頼性の高い判別器を得ることができるか．もう一つは，高次元のデータ空間において，いわゆる「次元の呪い」を以下に回避するか，そのための一般的な手法としてどのようなものがあるか，という問題である．これらの問題に対し，本論文は，次のような研究結果を報告した．

1. 訓練サンプル数が充分に得られない問題設定において，判別器のテスト性能の分散を考慮することにより，性能の悪い判別器が得られるリスクを回避したモデル選択を行う手法を提案した．そして，提案手法が，サンプルの出現の偏りにによる性能低下の影響を受けにくいモデル選択基準として有効であることを示した

2. 学習対象のデータ空間の次元が極めて高い場合，いわゆる「次元の呪い」の問題がある．近年，その原因として，高次元のデータ空間には，他の多くのデータとの類似性が高い，いわゆる，ハブとなるデータの存在が指摘されている．本論文では，ハブの存在を抑えることのできるカーネルとしてラプラシアンカーネルを取り上げ，それがハブの存在を抑制する性質を持つこと，および，それによりデータの分類性能の向上が得られることを実証した．他に，同様の性質を持つカーネルについても調査し，本論文での提案に対し，一般的な説明が可能であることを示した．

統計的機械学習における，サンプルの不足，および，高次元データについての問題への一般的な対処方法について議論し，新たな手法を提案した本研究は，独創性が高く，しかも実用面でも有用であり，大規模データ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．