論文内容の要旨

博士論文題目　Latent Variable Models for Bag-of-Words Data based on Kernel Embeddings of distributions
(分布のカーネル埋め込みに基づくBag-of-Wordsデータのための潜在変数モデル)

氏名　吉川友也

(論文内容の要旨)

機械学習や自然言語処理等の関連分野において，カーネル法は非線形予測を行うために研究されている．本論文では，入力データが特徴の多重集合，すなわちBag-of-Words（BoW）として表現される場合を考える．多くの論文では，予測性能の観点から，カーネル法は線形予測モデルよりも優れていることが報告されている．しかしながら，ガウシアンカーネルや多項式カーネル等の内積に基づくカーネル関数は，関連ある特徴間の相関をカーネルの計算に反映できないという共通の問題点がある．

この問題点を解決するために，BoWデータのためのカーネル法の新しい枠組みを提案する．この枠組みは（1）Latent Distribution Kernel（LDK）と呼ぶBoWデータのための潜在変数を含むカーネル関数のクラスの定義，(2) LDKに基づくモデルとそれらの最適化法の構築から構成される．

LDKでは，各特徴が低次元潜在ベクトルで表現されると仮定し，入力データはそのデータに含まれる特徴を表す潜在ベクトルの集合（分布）によって表現される．その分布をノンパラメトリックかつ効率的に表現するために，分布のカーネル埋め込みを用いる．この方法を用いることで，LDKは潜在ベクトルのすべての情報をデータ間のカーネルの計算のために利用するとともに，内積に基づくカーネル関数の問題点を解決することができる．

LDKの有効性を確認するために，機械学習における基本的な問題である分類，回帰，異種データ間マッチングに対して，LDKを組み込んだモデルを提案し，それらの最適化法を導出する．

実験では，各問題について，既存の線形や非線形手法と比較して，LDKを内包する提案モデルは予測性能が優れていることを示す．また，特徴の潜在ベクトルを可視化することにより，提案モデルの定量的有効性を示す．

氏　名	吉川友也

（論文審査結果の要旨）

平成27年7月24日に開催した公聴会の結果を参考に平成27年9月4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．

吉川友也は，本博士論文において，多重集合を素性とする分類問題や回帰問題に対して，Latent Distribution Kernelと呼ぶ新しい枠組みを提案し，従来の多次元素性空間を対象とした分類・回帰手法に比較して，有意な手法を提案した．本提案手法の特徴，新規性，有用性は以下のようにまとめることができる．

1. 自然言語処理などの応用研究では，単語を基本素性として用いることが多く，極めて高次元の事例を扱うことが必要になるだけでなく，単純な手法では，素性間すなわし単語間の関係（意味的な類似性）などの情報を扱うことが容易ではなかった．本提案では，各単語を低次元の潜在ベクトル空間内の点によって表現し，表現自体の学習が可能なモデルを提案した．

2. 問題の各入力事例を潜在ベクトルの分布によって表現し，その分布を効率的に扱うため，分布のカーネル埋め込みを利用することを提案した．そのため，入力データ間の関係を高次元空間で定義し，カーネル関数の計算により，潜在ベクトルのもつ情報を有効に利用しつつ，効率的な学習が可能であることを示した．

3. 分類，回帰，異種データ間のマッチングという機械学習における基本的な問題設定において，提案手法の適応実験を行い，従来手法を有意に上回る性能が実現可能であることを示した．

分布のカーネル埋め込みを利用し，潜在ベクトル空間における素性表現の学習を可能にした本研究は，独創性が高く，しかも実用的であり，機械学習および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．