OUKA
The University of Osaka
Institutional Knowledge Archive

Title	Towards Advanced Unsupervised Learning
Author(s)	Zhang, Bingyuan
Citation	大阪大学, 2023, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/92903
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

Form 3

Abstract of Thesis

Name （ZHANG BINGYUAN）	
Title	Towards Advanced Unsupervised Learning (高度な教師なし学習に向けて)

Abstract of Thesis

Unsupervised learning is an important category in machine learning algorithms. Unlike supervised learning, unsupervised learning is trained on unlabeled data without any explicit supervision, such as human-labeled data. This makes unsupervised learning an attractive option since labeled data can be expensive to obtain, while unlabeled data can be easily collected from a variety of sources. Unsupervised learning provides a powerful tool for exploring unlabeled datasets and extracting valuable information and insights. In fields such as computer vision, natural language processing, and biology, large-scale and high-dimensional datasets can be easily collected. Therefore, there is a growing demand for more efficient and high-performance unsupervised learning algorithms that can handle the complexity of large-scale and high-dimensional datasets.

In the first part, we study convex clustering. Clustering is a popular unsupervised learning technique. There are many famous clustering methods such as k-means and hierarchical clustering methods. Convex clustering is a modern clustering framework with both features of k-means and hierarchical clustering. Convex clustering is performed via solving a general fused lasso problem, which has a reputation of being difficult to be solved. We proposed a highly efficient L1-convex clustering method that can visualize the clusterpath on large datasets. Results show the proposed method is significantly more efficient than existing methods in finding the optimal solution.

In the second part, we study the hypothesis test on conditional independence. Conditional independence (CI) test is a fundamental problem in statistics that plays an important role in independence-based causal discovery. It aims to determine whether variables X and Y are conditionally independent given another variable Z fixed. However, when the conditioning set Z is continuous and high-dimensional, CI test becomes a challenging problem. We proposed a robust CI test that outperforms existing tests against the growth of dimension of Z. Results show our method provides a more accurate and reliable way of determining conditional independence.

様式７

論文審査の結果の要旨及び担当者

氏　　名　（ZHANG BINGYUAN）		
	（職）	氏　　名
論文審査担当者		
	主 査　教 授	鈴木 譲
	副 査　教 授	狩野 裕
	副 査　教 授	内田 雅之
	副 査　准教授	寺田 吉壱

論文審査の結果の要旨

本論文は、教師なし学習に関する一連のZhang Bingyuan氏の研究をまとめたものである。教師なし学習は、ラベル付けされていないデータを用いて訓練を行う統計的機械学習の一分野である。また、その特性から、大規模かつ高次元のデータセットを効率的に処理する手段として、コンピュータビジョン、自然言語処理、生物学等の領域での応用が期待できる。本論文の中でZhang氏は、クラスタリングおよび条件付き独立性の検定に関しての成果を主張している。

クラスタリングは、目的関数を定義してそれを最適にするデータ空間の分割を見出す理論と方法をさす。Zhang氏は、凸クラスタリングとFused Lasso(スパース推定の方法のひとつ)との類似性に着目して、最適解を最も高速な方法で求める方法を提案した。そのために、Johnsonの動的計画法によるアルゴリズムが該当問題に適用できることを明らかにし、サンプル数の線形時間でクラスタリングを実行できることを確認した。現在の結果は、目的関数がL1ノルムであることは前提とするが、さらなる高速化が期待できるとしている。

また、確率変数X,Y,Zに対して、ZのもとでのX,Yの条件付き独立性の検定の問題は、並べ替え検定が使えないため、通常のX,Yの独立性の検定の問題ほど容易ではない。また、サンプルが有限個しかないので、Zが連続で高次元であれば、格段に困難な問題となる。Zhang氏は、この問題に対して、k-meansクラスタリングによってZの領域を分割して、各領域で独立性検定に有用なHSIC(Hilbert Schdmit Information Criterion)を適用する方法を提案した。そして、Zの領域が高次元の場合に有用であることを、実験的に確認している。

これらの結果は、教師なし学習の分野において大きな貢献を果たしていると評価できる。また、博士論文公聴会におけるプレゼンテーションも明瞭であって、理路整然としていた。質疑の受け答えも的確であった。よって、本論文は博士（工学）の学位論文として価値のあるものと認める。