Form 3

# Abstract of Thesis

| Name （GUAN XIN） | |
|---|---|
| Title | On *k*-means Clustering for Complex High-Dimensional Data<br>(複雑な高次元データに対する*k*平均法クラスタリングについて) |


Abstract of Thesis

The *k*-means clustering is one of the most popular clustering methods, whereas it cannot perform well or even be inapplicable for data with non-linear clusters structure or with missing values, which is common in practice especially when the dimension of data is high. Existing methods like kernel *k*-means and *k*-POD clustering have been proposed for such complex cases. However, both of them are ineffective for high-dimensional data, due to the existence of noise features that have no contribution to the underlying clustering structure. Therefore, the purpose of this thesis is to make the *k*-means-based clustering effective for high-dimensional data with non-linear cluster structure and missing values.

The first contribution is to propose the sparse kernel *k*-means clustering for high-dimensional data with non-linear cluster structure. It assigns each feature a binary indicator and conducts the kernel *k*-means clustering while penalizing the sum of the indicators. The proposed method extends the advantages of kernel *k*-means clustering that can capture the non-linear cluster structure to the high-dimensional cases. An alternative minimization algorithm is proposed to estimate both the cluster labels and the feature indicators. We also prove the consistency of both clustering and feature selection of the proposed method.

The second contribution is to propose the regularized *k*-POD clustering for high-dimensional missing data. It introduces a regularization function of cluster centers to *k*-POD clustering, which shrinks cluster centers feature-wisely. The proposed method can reduce the bias of estimated cluster centers and improve clustering performance, for the high-dimensional missing data, where noise features that have no contribution to cluster structure are common. In addition, we propose a general framework of optimization based on the majorization-minimization algorithm, which has convergence guarantee.

The experiments on synthetic datasets and applications on real-world datasets verify the effectiveness and show better performance of the proposed methods. As a consequence, we extend the application of traditional *k*-means clustering to more complex data in the big data age.

様式７

# 論文審査の結果の要旨及び担当者

| 氏　名　（GUAN XIN） | | |
|---|---|---|
| | （職） | 氏　名 |
| 論文審査担当者 | 主 査　教 授 | 鈴木　譲 |
| | 副 査　教 授 | 杉本　知之 |
| | 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

本博士論文は、高次元データにおける k-means クラスタリングの適用を拡張し、非線形クラスタ構造および欠損値を持つデータに対して効果的な手法を提案している。従来の k-means 法は、データの高次元性や非線形クラスタ構造、欠損値の存在により適用が難しい場合が多く、これを克服するためにカーネル k-means や k-POD クラスタリングなどが提案されてきた。しかし、これらの手法も高次元データに対しては、クラスタ構造に寄与しないノイズ特徴の影響により有効に機能しないという課題があった。本論文では、この問題に対処するため、以下の二つの手法を提案し、その有効性を理論的および実験的に検証した。

第一の研究では、**スパース・カーネル k-means クラスタリング**を提案した。この手法では、各特徴量にバイナリの指標を導入し、カーネル k-means クラスタリングを行う際にその指標の総和をペナルティとして課すことで、クラスタ構造に寄与しない特徴量を自動的に選択する。本手法は、非線形なクラスタ構造を捉える能力を持つカーネル k-means の利点を維持しつつ、高次元データにも適用可能とするものである。また、クラスタラベルと特徴選択の指標を同時に推定するための交互最適化アルゴリズムを提案し、本手法の一致性（consistency）を理論的に証明した。

第二の研究では、**正則化 k-POD クラスタリング**を提案した。この手法は、高次元の欠損データにおける k-POD クラスタリングの性能向上を目的とし、クラスタ中心に対する正則化項を導入することで、特徴ごとのクラスタ中心の収縮を行う。このアプローチにより、ノイズ特徴量の影響を抑え、より適切なクラスタリングを実現した。さらに、マジョライゼーション・ミニマイゼーション（Majorization-Minimization）アルゴリズムを活用した一般的な最適化フレームワークを提案し、その収束性を保証した。

提案手法の有効性は、人工データセットおよび実データセットを用いた実験により検証され、従来手法と比較して優れたクラスタリング性能を示した。これにより、本論文は、従来の k-means クラスタリングの適用範囲を、ビッグデータ時代のより複雑なデータへと拡張することに成功している。

以上のように顕著な業績をあげており、本論文は、博士（理学）の学位論文として価値のあるものと認める。