様式３

論　文　内　容　の　要　旨

氏　名　（　寺　田　吉　壱　）	
論文題名	Statistical Theory of Clustering Methods for Multivariate Data （多変量データに対するクラスタリング法の統計理論）

論文内容の要旨

クラスタリング法は，教師無し学習の代表的な方法として，近年のデータの大規模化に伴いますますその重要性を増している．しかし，同時に多くの方法について統計的性質は明らかにされていない．これは実データ解析において非常に大きな問題となる．データの増加に伴い何らかの結果に収束しない方法では，同一の分布から生成した２つのデータに対してその方法を適用しても全く異なる結果が得られてしまい分析結果の解釈は本来行うことはできない．このような危険性があるにもかかわらず，統計的性質が議論されていない方法が多用されているのが現状である．そこで，本論文では，クラスタリング法の統計的性質として最も重要な性質である一致性に焦点をあてた．

多くの場合，データはある分布からランダム発生していると仮定する．この場合，データに対して最適なクラスタリングを得ることではなく，背後にある未知の分布に対して最適な結果を得ることが分析の目的となる．独立同一分布サンプリングの下では，サンプルサイズが大きくなると，データのもつ分布の情報が増えるため，クラスタリング結果も背後の分布に対して最適な結果が得られる事が期待される．大標本理論の枠組みでのクラスタリング法の研究では，サンプルサイズが無限大に発散すれば，背後の分布に対して最適な結果に収束するという一致性が重要となる．そこで，本論文では，次元縮約とクラスタリング法を同時におこなうReduced k-means（RKM）法とFactorial k-means（FKM）法に着目し，パラメータ空間のコンパクト性を仮定しない一般的な前提の下で，上述の一致性を証明した．これにより，クラスタリングの安定性に基づくクラスタ数と次元数の選択法を用いる事が出来る．さらに，一致性を通して，ある条件の下では，FKM法とRKM法が漸近同等となることを明らかにした．

近年，遺伝子データやfMRIデータのようにサンプルサイズよりも次元数（変量数）が大きいデータが多く得られるようになった．このようなデータに対しては，上述のような大標本理論の枠組みでの理論は当てはまらない．一方で，高次元データでは，ある程度の割合でクラスタ構造を反映する変量が含まれていれば，各サンプルに対するクラスタ情報が多いと考えられる．そのため，直感的には，全体の変量の増加に伴いクラスタ構造を反映した変量が増えれば，各サンプルのクラスタラベルを完全に特定できる完璧なクラスタリングが可能である．そこで，本論文では，サンプルサイズを固定し，次元数が発散するという高次元小標本理論の枠組みでは，距離の近さではなく距離の値にクラスタ構造が反映されている事を指摘し，非常にシンプルなクラスタリング法を提案した．具体的には，各サンプルの距離ベクトル間の差に基づき，従来のクラスタリング法を行う方法を提案し，提案手法が高次元小標本理論の枠組みで完璧なクラスタリング，すなわち，クラスタラベルの一致推定が可能であることを証明した．

様式7

論文審査の結果の要旨及び担当者

氏　　名　　（寺田　吉壱）		
論文審査担当者	（職）	氏　　名
	主 査　教 授	狩野　裕
	副 査　教 授	下平英寿
	副 査　教 授	内田雅之

論文審査の結果の要旨

学位申請者の研究業績は二つに大別される．一つ目は，クラスタリング法の数学的基礎付けの研究である．クラスタリング法は多変量データに基づき個体をいくつかの集団に分類する方法で，機械学習の分野では教師なし学習と呼ばれる手法の代表格である．クラスタリング法は記述的な多変量解析とみなされることが多く，歴史を刻んでいるにもかかわらず統計的推測の基礎付けは十分ではなかった．申請者は，reduced k-means (RKM) 法とfactorial k-means (FKM) 法に焦点を当て，母数空間のコンパクト性を課すことなく一般的な状況の下で，RKMとFKMの一致性（consistency）を証明することに成功した．また，RKMとFKMとはその定義から双対的な関係にあると言ってよいが，それらの類似性・相違性に関する数学的な研究は皆無であった．そこで，申請者はそれらが漸近同等になるための数学的条件を提出し，その条件が極めて厳しいことを指摘した．この結果はRKMとFKMとは基本的に異なるクラスタリング法であることを示している．これら研究成果は2編の英文学術論文として出版されることが決まっており，各所で高い評価を得ている．

二つ目の研究成果は高次元データに基づくクラスタリング法の開発である．近年ITやICT，そして各種のセンサリング技術の発展によって超高次元データが容易に採取できるようになったことから，それらを分析するための統計的方法論の開発が焦眉の急となっている．申請者は，高次元ほどクラスタリングの精度が向上する可能性があることに気づき，サンプルサイズを固定し次元数が発散するという高次元小標本理論の枠組みにおいて，クラスタラベルを一致推定できることを数学的に証明した．一般に高次元データの統計解析は困難を伴うことが多いが，高次元を味方に付けそれを有意義に利用することができることを示した価値は高い．

以上より，提出された論文は，博士（理学）の学位論文として価値があるものと認める．