氏名（本籍）	天野（あまの）　晃（こう）（福岡県）
学位の種類	博士（情報学）
学位記番号	博乙第2516号
学位授与年月日	平成22年6月30日
学位授与の要件	学位規則第4条第2項該当
審査研究科	図書館情報メディア研究科
学位論文題目	***k*-means クラスタリング法を適切に使用するためのシステム**

主	査	筑波大学教授	博士（学術）	中山伸一
副	査	筑波大学准教授	博士（情報科学）	真栄城哲也
副	査	筑波大学教授（連携大学院）	学術博士	深海薫
副	査	筑波大学名誉教授	理学博士	小野寺夏生
副	査	長浜バイオ大学教授	理学博士	池村淑道

論文の内容の要旨

学習型 *k*-means クラスタリング法は、他の非階層クラスタリング法に比べて計算時間が短くなるという特性を持つが、クラスタリング結果の初期値依存性が高いため、利用範囲が限定されていた。著者はその問題の解消に取り組み、1）初期値をランダムに設定せず、いくつかの初期配置状態を準備して状況に応じてそれを使う、2）クラスタ数を固定せず、学習過程において動的にクラスタを生成・統合する、という方法を考案した。そして、それらの機能を持った学習型 *k*-means クラスタリングを行うシステムとして Self-organizing clustering（SOC）を開発し、それを遺伝子のクラスタリングやシグナル発見、自然科学分野の雑誌のクラスタリング等、大規模な非階層クラスタリングを必要とする様々な問題に適用して提案する方法の妥当性を検証している。

本論文は6章から構成されている。

第1章では、複数の学習型非階層クラスタリング法の比較を行い、学習型 *k*-means クラスタリング法の他の方法に対する優位性と、初期値依存性という問題点を明らかにしている。それを受けて第2章では、初期値依存性について検討を行い、1）重心配置、周縁配置、対角配置、座標軸配置、格子状配置という初期配置状態の提案、2）初期配置状態の偏りを軽減する仕組みとしてのクラスタの動的な変化の提案、さらに3）適切な距離の選択を提案している。さらに SOC の実装を、初期配置状態設定機能と学習機能に分けて行い、2種類のコマンドを開発したこと、および GUI の実装等について説明している。

第3章では、SOC を用いた検討として、まず古細菌、真性細菌、真核生物という3つのドメインに対する DNA 配列のクラスタリング実験について述べている。3つのドメインのいずれかに含まれる629個の塩基配列データから得たオリゴヌクレオチド頻度ベクトルを用いて、ユークリッド距離と対角配置の組み合わせで検討を行った結果、3つのドメインにほぼうまく対応する DNA 配列のクラスタが得られることを明らかにしている。さらに明確な外部指標の無い nifH 遺伝子についての実験についても述べている。145個の nifH 遺伝子から得たオリゴヌクレオチド頻度ベクトルを用いて、ユークリッド距離と重心配置によるクラスタリングの結果、アラインメントに基づくクラスタリング結果と異なり生物種による遺伝子のクラスタが得られ

ることを明らかにしている。

第4章では、DNA配列中に出現するシグナル候補を検出する方法としてSOCを用いたFrame-cluster mapping（FCM）を提案し、線虫の全ゲノム塩基配列データ（約1億個程の塩基からなる）に適用した実験について述べている。FCMはゲノム塩基配列をシグナル候補を内包する長さの塩基配列に分断し、それをクラスタリングしてクラスタ毎にゲノム上の位置にマッピングすることにより、シグナル候補の探索を行う。線虫の全ゲノム塩基配列データを千個の塩基からなる断片（約10万個）に切断してそれらのオリゴヌクレオチド頻度ベクトルを用い、ユークリッド距離と重心配置によるSOCを用いたFCMにより、線虫のテロメア配列であるTTAGGCリピート領域と一致する領域を抽出できる事を明らかにしている。さらに従来の学習型*k*-meansクラスタリング法を用いたFCMを行い、同様の結果が得られないことを示している。

第5章では、被引用比率ベクトルに基づく自然科学分野の雑誌のクラスタリングにSOCを用いた実験について述べている。Journal Citation Reports 2004 Science Editionに含まれる5962誌について、コサイン距離と座標軸配置によるクラスタリングの結果、22個のクラスタが得られ、それらは既存の学術分野に対応することを明らかにしている。クラスタのサイズが比較的均等、クラスタの分離が良い、個別の雑誌の所属が妥当などの理由からSOCを用いたクラスタリングが適切であることを明らかにするとともに、従来の学習型*k*-meansクラスタリング法との比較を行い、クラスタの安定性、クラスタの分離性においてSOCの方が優れていることを示している。

第6章では、これらの実験結果を踏まえて、SOCの有用性を主張すると同時に、今後の課題として並列化が必要であることを述べている。さらにSOCの応用領域としてwebページのクラスタリング、文章のクラスタリング、メタゲノム解析などをとりあげてその可能性を考察している。

審査の結果の要旨

本論文の中核となるのは、学習型*k*-meansクラスタリング法の問題点を解消する方法論の提案と、その方法論がゲノム解析や引用分析等多方面に有効に利用できるという実証事例の紹介である。

学習型*k*-meansクラスタリング法は、*n*個の個体を*k*個のクラスタに分類する方法の1つである。基本的な考え方は、全ての個体とクラスタの組み合わせを発生させ、各クラスタ中心とそのクラスタに属する個体の距離の総和が最小になるとき、最適なクラスタリングが行われたと考える*k*-meansクラスタリング法にある。個体数が増えると全ての組み合わせを調べるのが困難であることから、個体を*k*個のクラスタにランダムに初期配置した状態から、機械学習によって解を求める学習型*k*-meansクラスタリング法が生まれた。機械学習は局所解を求めるものであるため、初期配置の状態に結果が大きく依存する。そのため、学習型*k*-meansクラスタリング法では、ランダムな初期状態を多数発生させ、それらの全てについて解を求めて、その中から最適なクラスタリング結果を選択する。

本論文では、クラスタリング対象となる個体群の持つ特徴ベクトルの分布状態をあらかじめ検討し、典型的なタイプのものについては最適な初期配置状態と距離の組み合わせを与えることを考案し、それを実現したシステムとしてSelf-organizing clustering（SOC）という環境を開発している。SOCは理論的な側面やアルゴリズムの革新性という意味での高い評価はできないが、後述する実証事例等を勘案すると、実際的な側面では意義が大きいといえよう。

DNA配列データを用いた実証事例として、古細菌、真性細菌、真核生物という3つのドメインに含まれるDNA配列のクラスタリング、明確な外部指標の無いnifH遺伝子のクラスタリング、線虫の全ゲノム塩基配列データからのシグナル候補の検出がある。前二者はSOCの結果を直接解釈するもので、生物種の同定を目的としている。このような目的の研究としては、Self-organizing Map（SOM）を用いたものがこれまで

行われて来たが、より明確にクラスタを示す事ができるSOCはSOMに対して有意性を持つものであり、SOCが妥当なクラスタリング結果を示したことは評価できる。最後のシグナル検出の事例では、Frame-cluster mapping（FCM）というシグナル候補の局在領域を検出する方法論もあわせて提案している。この方法は膨大な数の個体のクラスタリングを実施できなくては実現できないものであり、SOCの高速性のメリットが十分活かされている。線虫の全ゲノムを対象にして10万個もの断片のクラスタリングを行い、TTAGGCリピート領域と一致する領域を、その領域の特徴を前もって知ることなく抽出できたことは評価できる。

雑誌の引用データを用いた実証事例として、雑誌のクラスタリングがある。雑誌の学術分野による分類は、専門家による分野コードの付与などで実現しているが、その根拠に乏しい。被引用比率ベクトルを雑誌の特徴として雑誌をクラスタリングする方法は客観的な雑誌分類の方法論として評価できよう。なお、雑誌の引用データに基づき客観的に雑誌を分類する試みは他の非階層クラスタリング法を用いて同時期に行われており、この種の研究に対する注目度が高いことが伺える。

特徴ベクトルの次元数から見ると、DNA配列データの事例では2から5個の塩基からなるオリゴヌクレオチドの出現回数を特徴ベクトルとして捉えており、塩基の種類が4つあることから、16から1024次元のベクトルの扱いとなっている。また、雑誌の引用データの事例では5962誌の被引用数をもとにした被引用比率を特徴ベクトルとしており、5962次元ものベクトルを扱っている。実証事例で妥当なクラスタリングが行われていることは、SOCが少数次元はもとよりかなり多数次元の特徴ベクトルを持つ対象にも有効であることを示唆している。このように多様な次元の特徴ベクトルに対応できることを示したという点でも本論文の意義を評価できよう。

しかしながら、本論文の第6章にも一部述べられているが、初期配置状態の決定に個体の分布状態の知見が必要であるという問題が指摘される。分布状態をうまく反映できる配置を初期配置状態としたなら、安定な解に至る可能性は高く、また良い結果を得られるのは当然である。現状では、初期配置状態の決定は経験的ノウハウによって行っているように見受けられる。これをシステマティックに行える様なプロセスを考案すると、その実用性はより高まると考えられる。

総括すると、提案されている学習型*k*-meansクラスタリング法の問題点の解消法は、妥当なものであり一定の評価ができる。またSOCとして実装し、多様な事例に適用して各分野で新たな知見を得られたことは高く評価できる。

なお、SOCの全体像および3ドメインのDNA配列のクラスタリングについて書かれた本論文の核となる論文は、船井若手奨励賞を受賞しているが、その点も評価できよう。

よって、著者は博士（情報学）の学位を受けるに十分な資格を有するものと認める。