氏　　名	I Gde Putu Wirarama Wedashwara Wirawan （いぐで ぷとぅ ういららま うぇだすわら ういらわん）
授与学位	博　士(工学)
学位記番号	理工博甲第693号
学位授与年月日	平成28年３月17日
学位授与の要件	学位規則第４条１項
研究科，専攻の名称	理工学研究科(博士後期課程) 情報･デザイン工学系専攻
学位論文題目	Study on Evolutionary Data Mining for Database Clustering
論文審査委員	主　査　山口大学　教授　大林　正直 山口大学　教授　松藤　信哉 山口大学　教授　浜本　義彦 山口大学　教授　多田村　克己 山口大学　助教　間普　真吾

【学位論文内容の要旨】

Clustering is the unsupervised classification of patterns into groups. Database is one of the common inputs for clustering which is usually processed in the form of a matrix of attributes (features) and records (data). This dissertation is concerned with the optimization of database clustering using evolutionary computation and fuzzy database modeling, and proposes four rule based clustering algorithms. Rule based clustering is one of the solutions to provide automatic database clustering and interpretation of data patterns. Rule based clustering represents data patterns as rules by analyzing database structures on both of attributes and records. Each cluster is created by a rule pool where there are many rules which have similarity values to other rules in the internal cluster and dissimilarity values to the other rules in the external clusters.

The aim of the optimization proposed in this dissertation includes : improvement of clustering quality for large databases, making clusters with different capacity limitation, and online rule updating capability to handle data changes in databases. The optimization of database clustering is realized by evolutionary rule based clustering using genetic network programming (GNP).

GNP is an evolutionary optimization technique, which uses directed graph structures instead of strings in genetic algorithm or trees in genetic programming, which leads to enhancing the representation ability with compact programs derived from the re-usability of nodes in a graph structure. In this dissertation, GNP is used to handle rule extraction from databases by analyzing the attributes and records. Clustering is processed by grouping rules into the clusters based on the similarity measurement. This dissertation also uses fuzzy database modeling as a feature representation method to improve clustering ability to handle high dimensional databases. The proposed method optimizes the parameters of fuzzy membership

functions and automatically finds good fuzzy rules for making clusters.

This dissertation is composed of the following chapters.

Chapter 1 describes the background and objective of this study.

Chapter 2 discusses implementation of GNP and standard dynamic programming to realize a decision support system for record clustering in distributed databases, where the concept of knapsack problem (KP) is introduced to the cluster optimization. This chapter also discusses partial random rule extraction method in GNP to efficiently discover frequent patterns in a database for improving the clustering performance. The concept of KP in clustering is discussed, which is to distribute rules to each site by considering similarity (value) and data amount (weight) related to each rule to match the site capacities. From the simulation results using databases downloaded from UCI machine learning repository, it is clarified that the proposed method can create clusters considering the site capacities. In addition, to show the basic clustering ability of the proposed method, the clustering performance is compared to other five conventional clustering algorithms using six databases and the best average results are achieved.

Chapter 3 discusses database cluster optimization using GNP with online rule updating based clustering. In this chapter, online algorithm is utilized to maintain the cluster adaptability against several unbalanced data growth. To realize this ability, start node is added to represent the start positions of the node transition in GNP, and processing node which determines addition/deletion of rules and to which cluster each rule should be assigned is added to the conventional GNP based rule extraction method. The simulation results show that the better clustering results and iteration time comparing to GNP rule-based clustering without on-line adaptation.

Chapter 4 discusses a clustering method using GNP with the advantages of fuzzy object oriented database (FOOD) modeling. The main purpose of this chapter is to provide an additional mechanism to database clustering, that is, a data mining algorithm for extracting fuzzy rules, and building clusters based on the extracted fuzzy rules. The adoption of FOOD model to GNP rule extraction can increase the clustering quality and interpretability of clustering structures. The simulation results show the better clustering results comparing to GNP without FOOD and the conventional clustering methods.

Chapter 5 discusses a database clustering using GNP with feature selection and representation of fuzzy database. This chapter is an expansion of chapter 4 that aims for the improvement of clustering quality on high dimensional databases. Feature selection with fuzzy database is introduced in this chapter, where database is examined and many fuzzy membership functions are generated to calculate attribute relevancy. After the feature selection, attributes are grouped by considering their relevancy, then, rule extraction is performed by GNP separately using each group of attributes. The simulation results show that

the better clustering results are obtained for high dimensional databases comparing to the conventional clustering methods

Chapter 6 describes the conclusions of the dissertation and future research.

【論文審査結果の要旨】

コンピュータデータベースが開発された1960年代当初は，一つのデータベースを一台のコンピュータで利用するものであったが，1970年代に入り，ネットワークを経由して複数のコンピュータからアクセスされるようになった．その後，データ量が増大し，アクセス負荷が高まるにつれてデータベースの負荷を軽減するために分散型データベースが開発された．しかし，従来の分散型データベースは，設計者がデータの分散ルールを設定する必要があるため，その性能が設計者のスキルに大きく依存する問題や，データの傾向が変わったり，設計者が代わった場合などにメンテナンスが非常に困難である問題があった．また近年，ソーシャルネットワークサービスやオンラインストア，インターネットバンキングなど，膨大なデータの管理・通信が行われるようになるにつれ，データサーバの容量や負荷を考慮し，またユーザの要求に対して柔軟な対応ができるデータベースシステムを構築することが，効率的で利便性の高いネットワークサービスを実現する上で重要になっている．本論文では，分散型データベースの生成・管理の自動化を目的として，データマイニング，データクラスタリング，進化論的計算手法の概念を統合し，ルールに基づく新しいデータベースクラスタリング手法を提案し，その性能を検証している．

本論文のオリジナリティとして，

(i) 従来の，データ間の類似度に基づくクラスタリングではなく，データベースの性質を考慮したルール間の類似度に基づくクラスタリング手法を提案したこと，

(ii) 容量制限がある複数のサーバへのデータクラスタリングをナップサック最適化問題（資源配分問題）に対応させ，アンバランスな容量のサーバ構成にも対応可能なクラスタリング手法を実現したこと，

(iii) 時間変化に追随し，ルールの更新が可能な手法を実現したこと，

(iv) ファジイ理論を導入し，データ特徴量の判別の際，0/1の2値に判別するのではなく，中間的な値での判別も可能なクラスタリング手法を実現し，さらに特徴選択が可能な手法へ拡張したこと，

が挙げられる．

本論文の構成と内容は以下のとおりである．

第1章では，研究の背景と目的，および論文の構成について述べている．

第2章では，複数のデータサーバに対し，類似するデータを同一のサーバに保存する従来のデータベースクラスタリングを行うのみならず，サーバの容量が異なる場合にも適切なデータの分散が可能な手法を提案している．具体的には，遺伝的ネットワークプログラミング（GNP）を用いて頻出パターン（ルール）の抽出を行い，次に抽出ルールのクラスタリングをナップサック最適化問題に対応させて実現する手法を開発している．実データに対する評価実験の結果，提案手法は従来のクラスタリング手法と比較して優れたクラスタリング性能を持つことを確認している．

第3章では，データのパターンが変化した場合に，その変化を捉えオンラインでルールを更新する手法を提案している．第2章のルール抽出プロセスに，ルールの追加・削除・移動の機能を追加することで，オンラインルール更新なしの手法と比べて優れたクラスタリング性能をもつことを確認している．

第4章では，「安い」，「重い」などのあいまいなクエリにも対応できる分散型データベースとするため，ファジイ理論を導入したクラスタリング手法を提案し，ファジイを使用しない場合，および従来のクラスタリング手法と比較して優れた性能を持つことを確認している．

第5章では，多くの特徴量の中から重要なものを選別してルール抽出を行うことで，クラスタリングの性能を向上させることを目的とし，第4章のファジイクラスタリングによって得られる各特徴の重要度に基づく特徴選択法を提案している．評価実験の結果，提案手法は従来手法と比較して優れたクラスタリング性能を持つことを確認している．

第6章では、本論文の結論を述べている。

公聴会には、20 余名の参加者があり、活発な質疑応答がなされた。その主な質疑内容として、

(i) 従来手法のどの概念が提案手法のベースとなっているのか,

(ii) 第 2 章で，サーバの容量を考慮したクラスタリング手法を提案しているが，アクセス負荷を考慮したクラスタリングは可能か,

(iii) 第 3 章の比較実験が申請者の以前の提案手法との比較のみであるのはなぜか,

(iv) 提案手法はどのような応用システムを想定したものであるか，クラウドシステムへの応用は可能であるか,

(v) 本論文では，予め決められた数のクラスターにデータを分散させているが，クラスターの数が分からない場合の対応について解決法はあるか,

等の質問があり，いずれの質問に対しても申請者から的確な回答がなされた.

以上より，本研究は，新規性，信頼性，有効性，実用性ともに優れており，博士（工学）の論文として十分に値するものと判断した.

論文内容，審査会，及び公聴会での質問に対する応答などから総合的に判断して，最終試験は合格とした.

なお，査読のある関連論文の発表状況は以下の通りである.　（下記以外　国際会議会議録　4 編）

1. Wirarama Wedashwara, Shingo Mabu, Masanao Obayashi, Takashi Kuremoto, On-line Rule Updating System Using Evolutionary Computation for Managing Distributed Database, Journal of Robotics, Networks and Artificial Life, Vol. 2, No. 2, pp. 73—78,2015
2. Wirarama Wedashwara, Shingo Mabu, Masanao Obayashi, Takashi Kuremoto, Combination of Genetic Network Programming and Knapsack Problem to Support Record Clustering on Distributed Databases, Expert Systems with Applications, An International Journal, Vol. 46, pp. 15–23, 2016