HOKKAIDO UNIVERSITY

Title	A Study of Efficient and Robust Clustering for Large-Scale Datasets [an abstract of dissertation and a summary of dissertation review]
Author(s)	劉, 浩
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11521号
Issue Date	2014-09-25
Doc URL	https://hdl.handle.net/2115/57167
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Hao_Liu_abstract.pdf, 論文内容の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　劉 浩

学 位 論 文 題 名

A Study of Efficient and Robust Clustering for Large-Scale Datasets

（大規模データセットに対する効率的で信頼性の高いクラスタリングに関する研究）

Clustering, also called cluster analysis, is one of the most important techniques for analyzing data, and has been widely applied in many fields, such as data mining, machine learning, pattern recognition, information retrieval, etc. However, the amount of data is exponentially increasing because of the recent big data explosion. Many excellent clustering algorithms designed with good robustness, e.g. the recently proposed fuzzy density-based clustering algorithms, are not efficient enough for large-scale datasets. On the other hand, some efficient algorithms, e.g. the incremental clustering approaches, are not robust enough when they learn new data, especially in a noisy environment. In this dissertation, two novel methods for efficient and robust clustering are proposed. One is an efficient fuzzy density-based clustering algorithm which uses a new concept, called “ landmark ”, to obtain a good representation for the input dataset. Since the number of landmarks is much smaller than that of the actual data, the computational cost is significantly reduced. The other one is a robust incremental self-organizing neural network, where a new flexible network structure is modeled so that the algorithm is able to automatically determine a suitable number of neurons for the network while keeping stability to outliers.

In Chapter 1, the background of this study, including the purpose of clustering and a brief summarization of the state of the art, is introduced. Then the challenge of detecting clusters in large-scale datasets is discussed. After that, the purpose of this study and a brief introduction of the two proposed methods as well as a description of their contributions are presented.

Chapter 2 presents the fundamentals necessary for the remaining chapters. Three classical clustering techniques, i.e. partitional, hierarchical and density-based clusterings, are described and their benefits and drawbacks are discussed. Then it is followed by the discussion on incremental clustering, including classical ones and relatively new ones based on the self-organizing neural networks.

In Chapter 3, as the base of this study, a density-based clustering method, called Landmark Fuzzy Neighborhood DBSCAN (landmark FN-DBSCAN), is presented, aiming to improve the efficiency of FN-DBSCAN. The new concept, landmark, is introduced to represent a subset of the input dataset so that the original dataset can be compressed and represented by the generated landmarks. We give a

theoretical analysis on the time and space complexities of the algorithm, which shows that both of them are linear to the size of the dataset. The experiments presented in this chapter also show that the landmark FN-DBSCAN algorithm can be much faster than the FN-DBSCAN, while maintaining the good clustering quality of the original algorithm. However, we will see that this algorithm has a critical limitation, because the final clustering result may be easily influenced by the input data orders, if the number of landmarks is very small.

Chapter 4 introduces a new effective and efficient solution for the task of landmark generation, i.e. Locally Adaptive Incremental LBG (LAI-LBG). In order to overcome the above-mentioned drawback, LAI-LBG incrementally inserts and removes landmarks with no requirement of predefining the number of landmarks, and it is not sensitive to the initializations and the input data orders, which means that the algorithm can pursue a well-balanced representation of the input dataset with a small number of landmarks. In addition, it has a higher efficiency than its competitors with comparable quantization quality.

In Chapter 5, a novel density-based clustering algorithm which is a significant improvement of the landmark FN-DBSCAN algorithm is proposed. Benefiting from LAI-LBG, the new algorithm can obtain a good clustering result with a few number of landmarks, which indicates that the efficiency has further increased. In addition, it is not sensitive to the input data orders, which means that the robustness is improved. Furthermore, its parameters have typical values or can be automatically estimated, which means that the practicability is also enhanced.

Chapter 6 presents a novel incremental self-organizing neural network, called enHanced Incremental Growing Neural Gas (Hi-GNG). In this work, an enhancement of the conventional competitive Hebbian learning (enhanced CHL) is also proposed, where the birth and death of neurons, the creation and elimination of connections between each pair of neurons, and the adjustment of the connection-strength of each connection are modeled. On the basis of the enhanced CHL, the Hi-GNG algorithm can automatically generate a topological structure with a suitable number of neurons. Furthermore, it is also robust to noisy data and even more efficient than its competitors, such as the Growing Neural Gas (GNG) algorithm.

In Chapter 7, the methods proposed in this dissertation are summarized and some future works are discussed.