효율적 군집화를 위한 탐색 방법 연구

전진호*

*관동대학교

A Study of Search Methodology for Efficient Clustering

Jin-Ho Jeon*

*Kwan Dong University

E-mail : jhjy@kd.ac.kr

요 약

경제, 경영, 의료 및 공학 등 실세계의 많은 시스템들은 복잡한 현상을 갖는다. 이러한 특징의 시스템들을 이해하는 방법은 시스템에 대한 모델을 세우고 분석하는 것이다. 모델을 세우고 분석하는 과정은 두 단계로 이루어진다. 첫째, 데이터에 대하여 효율적 군집을 결정하는 과정, 둘째, 각 군집에 대한 적합한 모델을 생성하는 과정이다. 본 연구에서는 효율적 군집화를 위한 탐색 방법에 대하여 살펴본다.

ABSTRACT

Most real world system such as world economy, management, medical and engineering applications contain a series of complex phenomena. One of common methods to understand these system is to build a model and analyze the behavior of the system. As a first step, Determining the best clusters on data. As a second step, Determining the model of the cluster. In this paper, we investigated heuristic search methods for efficient clustering.

키워드

시계열데이터, 군집, 기준

Ⅰ. 서 론

실세계의 시스템들은 동적인 특징을 가지며 시간적인 특징들에 의해서 묘사되고, 그것들의 값들은 관측기간 동안 의미 있게 변한다. 이렇게 시간의 흐름에 따라 발생한 데이터를 수집하여 기록한 것을 시계열 데이터라 한다. 이러한 시계열 데이터를 분석하여 내포하고 있는 특징을 찾아낸다면 그러한 특징들은 시계열 데이터를 이해하고 분석하는데 도움을 줄 것이다.

시계열 데이터의 분석을 위한 과정은 두 단계로 구성된다. 첫 번째, 유사한 시계열 데이터들의 군집화 과정이다. 두 번째는 각 군집마다 적합한 모델을 학습하는 과정이다.

본 연구에서는, 첫 번째 단계에서 효율적 군집수를 결정하기 위한 휴리스틱 방법론을 살펴본다.

Ⅱ. 배경 연구

군집화는 클래스 정보가 기록되지 않은 데이터를 가정한다. 목적은 그룹 내에서는 데이터 유사도가 그리고 그룹들 사이에서는 데이터의 비유사도가 최적이 되도록 구조를 생성하는 과정이다.

시계열 데이터의 군집에 대한 연구는 크게 세 범주로 구분되어 진다. 첫째, 시계열 데이터 쌍의 객체 또는 거리측정을 이용하는 correlation measure[1]와 Longest common sequence measure[2], 그리고 Dynamic time warping[3]와 같은 근사기반 방법론이다. 둘째, 각 시계열 데이터들로부터 특색을 이루는 특징의 집합을 추출하여 이용하는 Fourier descriptor[4], 그리고

Wavlet analysis[4]과 같은 특징기반 방법론이다. 세 번째, 데이터에 가장 적합한 모델의 집합을 찾는 모델기반 방법론이다. 이에는 회귀모델, 은닉마코프모델[5]등이 있다.

은닉마코프모델은 시계열 특징으로 묘사되는 시계열 데이터의 표현 모델링에 적합하다. 이유는 각 상태에서 특징들에 대한 적합한 확률함수를 사용하여 연속적인 값을 갖는 시계열 데이터를 쉽게 처리하며, 다수의 시계열 특징을 가진 데이터의 묘사가 쉽기 때문이다.

Ⅲ. 효율적 군집 결정에 적용 가능한 탐색 방법

주어진 데이터 객체들의 효율적인 군집의 경우는 다양할 것이다. 최악의 경우 데이터의 수만큼 군집화될 수 있을 것이며, 이는 매우 큰 비용을 발생시킨다.

본 연구의 주된 관심은 시계열 데이터의 효율적 군집화를 결정할 수 있는 탐색 방법론을 살펴보는 것이다.

3.1 휴리스틱 기준 - Bayesian Information Criterion(BIC)

BIC는 Laplace approximation으로부터 유도된다.

$$\log P(M \mid X) \approx \log P(X \mid M, \hat{\theta}) - \frac{d}{2} \log N \qquad (1)$$

위 식(1)에서 d는 모델에서 파라미터의 수이다. N은 데이터 객체들의 수이고 $\hat{\theta}$는 모델 M의 한계우도(ML)의 파라메터 구성이다. 식에서 첫 번째 항은 데이터를 가장 잘 설명할 수 있는 상세한 데이터의 모델을 찾도록 유도하는 성분이다. 두 번째 항은 은 모델 내의 파라메터 개수에 대한 penalty 항으로 볼 수 있다.

3.2 휴리스틱 기준 - Cheeseman-Stutz Approximation(CS)

cheeseman-stutuz는 베이지안 클러스터링 시스템, AUTOCLASS[6]에서 제안되었다.

$$P(X \mid M) = P(X' \mid M) \frac{P(X \mid M)}{P(X' \mid M)} \qquad (2)$$

위 식(2)에서 첫 번째 항은, 데이터의 한계우도를 나타낸다.

두 번째 항은 조정항이다. 두 번째 항에서 BIC 측정을 적용하여 확장하면 다음 식(3)을 얻을 수 있다.

$$\log P(X \mid M) \approx \log P(\hat{\theta} \mid M) + \log P(X \mid \hat{\theta}, M) \qquad (3)$$

위 식(3)에서 X는 불충분한 데이터이다. $P(\hat{\theta} \mid M)$는 모델 파라메터들의 한계우도이다.

BIC[7]와 CS[6]는 이러한 두 항에 상호 배타적인 특성이 서로 조화되는 타협점에서 정확하지는 않지만 효율적인 군집화를 결정할 수가 있다.

이러한 탐색 방법론의 주된 아이디어는 주어진 기준 함수를 통해 하나의 군집으로부터 출발하여 군집 수를 하나씩 증가하여 가장 높은 기준함수의 값을 갖는 군집의 수가 효율적 군집으로 결정됨을 나타낼 수 있다.

Ⅳ. 결 론

실세계의 시스템들은 동적인 특징을 가지며 시간적인 특징들에 의해서 묘사된다. 이렇게 시간의 흐름에 따라 발생한 데이터를 수집하여 기록한 것을 시계열 데이터라 한다. 시계열 데이터의 분석을 위한 과정은 두 단계로 구성된다. 첫 번째, 군집화 과정이다. 두 번째는 각 군집마다 적합한 모델을 학습하는 과정이다.

본 연구에서는, 첫 번째 단계, 즉, 효율적 군집화를 위하여 적용될 수 있는 탐색 방법론에 대하여 BIC와 CS 방법론을 살펴보았다.

향 후 연구되어야 할 부분은 본 연구에서 살펴본 방법론들이 동적인 특징을 갖는 실세계의 시스템들에서 발생되어지는 시계열데이터들에 대하여 직접적인 적용을 통하여 시스템에 대한 일반적인 분석모델을 수립할 수 있는 부분으로 확대가 되어 질 수 있는지 확인을 하는 과정이 필요하다고 생각된다.

참고문헌

[1] A.K. Jain and D. C. Dube, Algorithem for clustering data, Prentice Hall, 1988.

[2] D. S. Hirschberg, "Algorithem for longest common subsequence problem," Journal of Association of Computer Machine 24, pp664-675,1977.

[3] T. Oates, " Identifying distinctive subsequences in multivariate time series by clustering," Proceedings of the sixteenth International Conference on Machine Learning, 1999.

[4] Y. Huhtala, J. Karkkinen, H. Toivonen, and N. R, "Mining for similarity in aligned time series using wavlets," Proceedings of SPIE on Data Mining and knowledge Discover: Theory, Tools, and Technology, 1999.

[5] L. Rabiner, " A tutorial on Hidden Markov Models and selected applications in

speech recognition," Proc. of IEEE77, pp.257-286, 1989.

[6] Cheeseman, P., and Stutz, J. "Bayesian classification(autoclass)"

[7] Heckerman, D., Geiger, D., and Chekering, D. M. "A tutorial on learning with bayesian networks," machine Learning 20, pp.197-243, 1995.