논문 2011-5-19

Temporal 데이터의 효율적 군집 추정을 위한 기준 연구

A Study of Criterion for Efficient Clustering Estimation of Temporal Data

전진호*, 김민수**

Jin-Ho Jeon, Min-Soo Kim

요 약 실세계에서 사용되는 많은 정보시스템들은 복잡한 동적 현상을 나타낸다. 이러한 동적 현상을 갖는 정보시스템들을 이해하는 방법은 시스템에서 발생된 데이터들을 통하여 모델을 세우고 분석하는 것으로서 동적 현상을 이해할 수 있다. 모델을 세우고 분석하는 과정은 두 단계로 이루어진다. 첫 번째는 시스템에서 발생되는 대용량의 데이터에 대하여 효율적 군집을 결정하는 과정이며, 두 번째 과정은 각 군집에 대한 적합한 모델을 결정하는 과정이다. 본 연구에서는 두 과정 중 첫 번째 과정인 대용량 temporal 데이터들에 대하여 정확한 군집 수를 추정하기 위한 기준들을 살펴보고 인공적으로 실험데이터를 생성하여 실험을 하였다. 실험 결과 살펴본 베이지안정보기준이 올바른 군집 수를 추정하는 결과를 갖는 것을 확인하였다.

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. It is also confirmed that the Bayesian Information Criterion more reliable than Cheeseman-Stutz ones.

Key Words : Temporal 데이터, 군집, 기준, 한계우도

Ⅰ. 서 론

실세계의 많은 정보시스템들은 동적인 특징을 가진다. 즉, 시간적인 특징들에 의해서 묘사되고, 그것들의 값들은 관측기간 동안 의미 있게 변함을 의미한다. 이렇게 시간의 흐름에 따라 발생한 데이터를 수집하여 기록한 것을 temporal 데이터라 한다.

temporal 데이터 내에 내재하는 속성들, 물품 또는 사건들을 통해 연관성 또는 순차 패턴과 같은 특징이 명확한 분야의 연구는 많이 진행되어왔다[1]. 이러한 temporal 데이터를 분석하여 내포하고 있는 특징을 찾아낸다면 그러한 특징들을 통하여 temporal 데이터를 이해하고 분석하는데 도움을 줄 것이다.

대용량의 temporal 데이터의 분석을 위한 과정은 두 단계로 살펴볼 수 있다. 첫 번째 과정은 발생되어진 대용량의 temporal 데이터들을 유사한 데이터 객체들끼리의 군집화 과정이다. 두 번째 과정은 각 군집을 잘 설명할 수 있는 적합한 모델을 생성, 학습하는 과정이다.

본 연구에서는, 위의 두 단계 과정 중에서 첫 번째 과정, 즉, temporal 데이터의 군집화 과정에서 데이터들의 특징을 표현하기에 적합한 군집 수를 추정하는 기준에

*정회원, 관동대학교, 경영학과
**정회원, 관동대학교, 호텔경영학과
접수일자 2011.6.30, 수정일자 2011.9.19
게재확정일자 2011.10.14

대하여 살펴보고 실제의 temporal 데이터, 즉 주식데이터를 통하여 모델을 결정하고 모델로부터 실험을 위한 인공적인 여러 형태의 데이터집합을 임의로 생성한 후 실험데이터에 맞는 적합한 군집 수를 추정하는지 각 기준들을 확인하였다.

II. 배경 연구

군집화의 목적은 그룹 내에서는 데이터 유사도가 그리고 그룹들 사이에서는 데이터의 비유사도가 최적이 되도록 구조를 생성하는 과정으로 temporal 데이터의 군집에 대한 연구는 크게 세 범주로 구분되어 진다. 첫째, 근사기반 방법론이다. 이는 temporal 데이터 쌍의 객체 또는 거리측정을 이용하는 correlation measure[2]와 longest common sequence measure[3], 그리고 dynamic time warping[4]등이 있다.

둘째, 특정기반 방법론이다. 이는 각 temporal 데이터들로부터 특색을 이루는 특징의 집합을 추출하여 이용하는 fourier변환, descriptor[5], 그리고 wavlet analysis[5]등이다.

세 번째, 모델기반 방법론이다. 이는 데이터에 가장 적합한 모델을 생성하여 모델간의 한계우도를 통하여 유사성을 측정하는 것이다. 이에는 회귀모델, 마아코프모델. 은닉마코프모델[6]등이 있다.

은닉마코프모델은 temporal 특징으로 묘사되는 temporal 데이터의 표현 모델링에 적합하다. 이유는 각 상태에서 특징들에 대한 적합한 확률함수를 사용하여 연속적인 값을 갖는 temporal 데이터를 쉽게 처리하며, 다수의 temporal 특징을 가진 데이터의 묘사가 쉽기 때문이다. 이에 따라 본 연구에서는 데이터에 대한 모델에 은닉마코프모델을 적용한다.

III. Temporal 데이터의 군집 수 추정을 위한 기준

주어진 데이터들에 대한 군집의 경우의 수는 다양할 것이다. 최악의 경우 데이터의 수만큼 군집화 될 수 있을 것이며, 이는 매우 큰 비용을 발생시킨다.

본 연구의 주된 관심은 temporal 데이터의 효율적인 군집 수를 결정하기 위한 기준들을 살펴보고 제시되는 방법들이 효율적인 군집 수를 추정하는지 확인하는 것이다.

주어진 데이터가 완전할 때, 즉 관측 값들이 모델의 모든 변수들에 대응되어지면 모델에서 모든 변수들은 독립성을 가정한다. 이러한 경우에서는 한계우도의 계산이 매우 간단하다. 그러나 주어진 데이터가 불완전할 때, 모델의 변수들에 대응되는 관측 값들이 없는 경우에는 모델에서 변수들 간에 종속적인 된다. 이러한 경우에서는 정확한 폐형해를 얻는 것은 매우 복잡하다. 그러므로 일반적으로 근사기법들이 사용된다. 한계우도를 구하기 위한 근사기법은 다음과 같다. 몬테카를로 기법, 라플라스 근사법 등이 있다. 그러나 이 방법들은 정확한 결과 값에 수렴하지만 계산이 복잡하여 일반적인 데이터 셋에서 실제 사용하기에는 비용이 너무 많이 드는 것으로 알려져 있다. 이 방법들 보다 정확도는 약간 떨어지지만 계산복잡도를 많이 줄여 효율성을 준 베이지안정보기준(BIC), Cheeseman-Stutz(CS) 근사 등이 있다[7].

1. Bayesian Information Criterion(BIC)

베이지안정보기준(BIC)은 라플라스 근사법(laplace approximation)으로부터 유도된다. 대량의 데이터에 대하여, $P(\theta|X,M) \propto P(X \mid \theta,M)\ P(\theta \mid M)$는 다변량 가우시안 분포로서 근사되어진다.

$P(X \mid \theta,M)\ P(\theta \mid M) \equiv e^{g(\theta)}$로 정의하면,

$$g(\theta) = \log\left(P(X \mid \theta,M)\ P(\theta \mid M)\right), \quad (1)$$

$g(\theta)$를 최대화시키는 θ를 $\hat{\theta}$로 정의하면, 이 매개변수의 조합은 $P(\theta|X,M)$를 최대화시키므로 이것을 최대사후확률(MAP)이라 부른다. $g(\theta)$에 대해 2차 테일러 다항식으로 근사시키면 $\hat{\theta}$에서 다음이 성립된다.

$$g(\theta) \approx g(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})\,A(\theta-\hat{\theta})^t, \quad (2)$$

여기서, $(\theta-\hat{\theta})^t$는 행벡터 $(\theta-\hat{\theta})$의 트랜스포즈이고 A는 $\hat{\theta}$ 조합에서 $g(\theta)$에 대한 음의 Hessian이다. $e^{g(\theta)}$과 식(1)를 이용하면 다음의 식(3)을 얻는다.

$$P(X \mid \theta,M)\ P(\theta \mid M) = e^{g(\theta)}$$

$$\approx e^{\left\{ g(\hat{\Theta}) - \frac{1}{2}(\Theta-\hat{\Theta})A(\Theta-\hat{\Theta})^t \right\}}$$

$$= e^{g(\hat{\Theta})} e^{-\frac{1}{2}(\Theta-\hat{\Theta})A(\Theta-\hat{\Theta})^t}$$

$$= P(X \mid \hat{\theta}, M)\ P(\hat{\theta} \mid M)\ e^{-\frac{1}{2}(\theta-\hat{\theta})A(\theta-\hat{\theta})^t} \quad (3)$$

식(3)을 Θ에 대하여 적분하면, 다음과 같이 전개된다.

$$\int P(X \mid \theta, M)\ P(\theta \mid M)\, d\theta = \int P(X, \theta \mid M) d\theta = P(X \mid M)$$

$$\int P(X \mid \hat{\theta}, M)\ P(\hat{\theta} \mid M)\ e^{-\frac{1}{2}(\theta-\hat{\theta})A(\theta-\hat{\theta})^t} d\theta$$

$$= P(X \mid \hat{\theta}, M)\ P(\hat{\theta} \mid M) \cdot (2\pi)^{\frac{d}{2}} / \sqrt{|A|}$$

$$P(X \mid M) = P(X \mid \hat{\theta}, M)\ P(\hat{\theta} \mid M) \cdot (2\pi)^{\frac{d}{2}} / \sqrt{|A|} \quad (4)$$

(4)의 양변에 로그를 취하면

$$\log P(X|M) \approx \log P(X|\hat{\theta}, M) + \log P(\hat{\theta}|M) + \frac{d}{2}\log(2\pi) - \frac{1}{2}log|A| \quad (5)$$

$$\log P(M \mid X) \approx \log P(X \mid M, \hat{\theta}) - \frac{d}{2} log N \quad (6)$$

위 식(6)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 패널티(penalty) 항으로 볼 수 있다[8].

2. Cheeseman-Stutz(CS) Approximation

cheeseman-stutuz는 베이지안 클러스터링 시스템, AUTOCLASS[8]에서 제안되었다.

$$P(X \mid M) = P(X' \mid M)\frac{P(X \mid M)}{P(X' \mid M)} \quad (7)$$

위 식(7)에서 첫 번째 항은, 데이터의 한계우도를 나타낸다.

두 번째 항은 조정항이다. 두 번째 항에서 BIC 측정을 적용하여 확장하면 다음 식(8)을 얻을 수 있다.

$$\log P(X \mid M) \approx \log P(\hat{\theta} \mid M) + \log P(X \mid \hat{\theta}, M) \quad (8)$$

위 식(8)에서 X는 불충분한 데이터이다. $P(\hat{\theta} \mid M)$는 모델 파라메터들의 한계우도이다.

베이지안정보기준[10]과 Cheeseman-StutzS[9]는 이러한 두 항에 상호 배타적인 특성이 서로 조화되는 타협점에서 정확하지는 않지만 효율적인 군집 수가 결정된다.

이러한 휴리스틱 기준 방법론의 주된 아이디어는 주어진 기준 함수를 통해 하나의 군집으로부터 출발하여 군집 수를 하나씩 증가하여 가장 높은 기준함수의 값을 갖는 군집의 수가 효율적 군집 수로 결정됨을 나타낸다.

Ⅳ. 실 험

군집 수 추정을 위한 판단기준으로 살펴 본 베이지안 정보기준(BIC)과 Cheeseman-Stutz(CS)의 효용성을 실험을 통하여 확인한다.

실험을 위하여 사용된 데이터는 2005년도 5월3일부터 12월5일로서 데이터의 길이가 150일인 전기전자, 유통, 제조업의 실제 주가 데이터를 이용하였다. 세 업종의 데이터를 통해 각 업종에 대한 모델을 생성한 후, 생성된 모델들로부터 랜덤하게 시계열데이터를 생성하였다.

실험의 방법은 두 업종의 데이터를 통한 군집 추정과 세 업종의 데이터를 통한 군집 추정으로 나누어 실행하였다. 각각의 두 경우 모두 올바르게 추정하는지 확인한다. 각각의 경우에 대하여 데이터의 길이에 따른 실험과 데이터 객체 수에 따른 실험을 하여 군집 수 추정에 있어 길이와 객체 수에 따라 결과의 차이를 확인하였다. 데이터의 길이에 따른 실험을 위하여 데이터의 길이는 30일과 60일의 데이터로 생성하였으며 데이터의 객체 수에 따른 실험을 위해서는 각 업종별로 3개, 5개, 7개의 군집으로 구성하였다.

1. 두 업종의 데이터를 통한 추정

두 업종에서 군집 추정을 위한 실험으로 전기전자와 제조업의 데이터를 사용하였다. 먼저 데이터 객체의 길이에 따라 군집 수를 정확하게 추정하는지 살펴보기 위하여 두 업종의 모델로부터 생성된 데이터의 길이를 30

일과 60일로 하였으며 각 모델별로 데이터 객체는 5개로 하였다.

그림 1, 그림 2, 그림 3 그리고 그림 4에서 X축은 추정된 군집 수를 Y축은 각 군집에서의 우도값을 나타낸다. 그림 1과 그림 2는 베이지안정보기준(BIC) 추정의 결과를 보여준다.

결과를 보면 데이터의 길이에 관계없이 30일, 60일 모두에서 군집 수를 2개의 군집으로 정확히 추정하는 것을 확인할 수 있다.

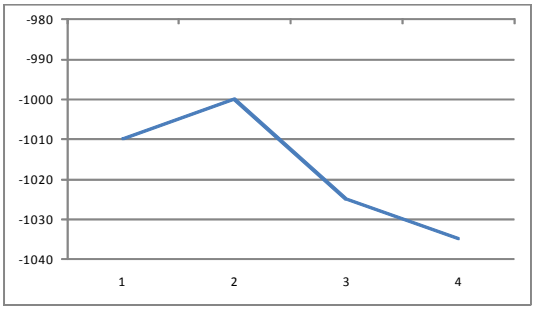

그림 1. 길이(30일)에 따른 BIC 추정
Fig. 1. BIC Estimation by data length(30days)

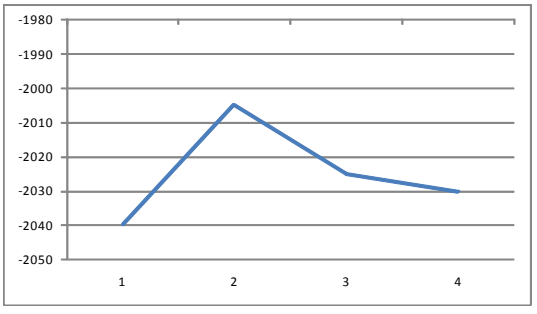

그림 2. 길이(60일)에 따른 BIC 추정
Fig. 2. BIC Estimation by data length(60days)

그림 3과 그림 4를 보면 데이터 객체의 길이에 따른 Cheeseman-Stutz(CS) 추정의 결과를 보여준다. 결과를 보면 데이터 객체의 길이가 30일인 경우에는 군집 수를 1개로 잘못된 추정을 보여주고 있으며, 데이터 객체의 길이가 60일인 경우에는 군집 수 2개를 정확히 추정하는 것을 확인할 수 있다.

즉 Cheeseman-Stutz(CS) 추정의 결과에서는 데이터 객체의 길이에 따라 정확히 추정이 이루어지지 않는 것을 확인할 수 있다.

데이터 객체의 길이에 따라 군집 수를 정확히 추정하는지를 살펴보는 실험에서는 Cheeseman-Stutz(CS) 근사기법보다 베이지안정보기준(BIC)의 방법론이 데이터 객체의 길이에 영향을 받지 않고 더 정확한 군집 수를 추정하는 것을 보여주었다.

두 번째는 데이터 객체의 길이는 60으로 동일하며 객체의 수를 다르게 한 경우이다. 각 모델별로 데이터 객체의 수를 3, 5 그리고 7개인 경우를 살펴보았다. 결과는 아래 표 1과 같다.

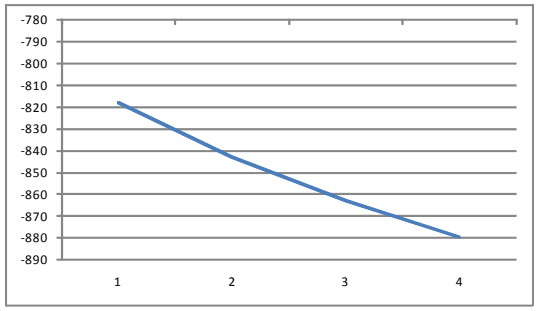

그림 3. 길이(30일)에 따른 CS 추정
Fig. 3. CS Estimation by data length(30days)

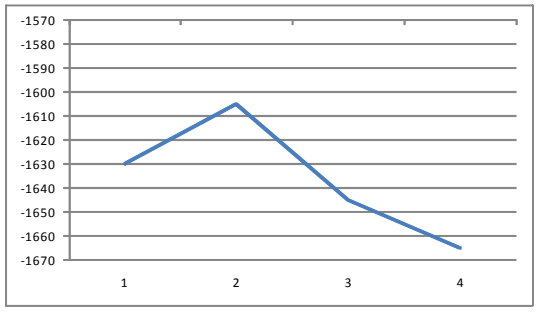

그림 4. 길이(60일)에 따른 CS 추정
Fig. 4. CS Estimation by data length(60days)

표 1. 데이터 객체 수에 따른 추정 결과
Table 1. Estimation result by data object number

베이지안정보기준 (BIC) 추정		Cheeseman-Stutz (CS)추정	
3 개체	2 군집	3 개체	1 군집
5 개체	2 군집	5 개체	2 군집
7 개체	2 군집	7 개체	2 군집

표 1을 보면 베이지안정보기준(BIC) 추정에서는 데이터 객체 수에 관계없이 군집 수를 2개 정확히 추정하는 것을 확인 할 수 있다. 반면 Cheeseman-Stutz(CS) 추정

에서는 데이터 객체가 5개와 7개인 경우에는 정확히 군집 수를 추정하나 3개인 경우에는 군집 수를 1개 추정함으로서 부정확하게 추정하는 것을 확인할 수 있다.

2. 세 업종의 데이터를 통한 추정

전기전자, 유통, 제조업의 세 업종 데이터를 통한 군집 추정에서도 앞에서의 실험과 같이 데이터의 길이와 객체 수에 따라 군집 수 추정에 적용하였다.

그림 5와 그림 6은 베이지안정보기준(BIC) 추정의 결과를 보여준다. 결과를 보면 데이터의 길이에 관계없이 30일, 60일 모두에서 군집 수를 3개의 군집으로 정확히 추정하는 것을 확인할 수 있다.

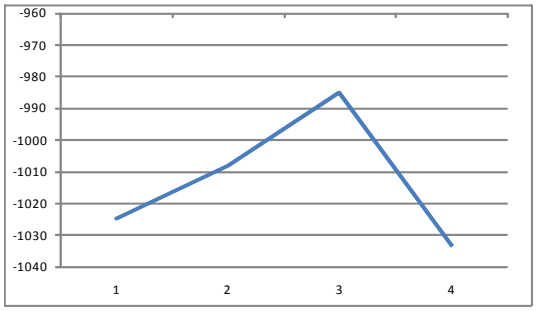

그림 5. 길이(30일)에 따른 BIC 추정
Fig. 5. BIC Estimation by data length(30days)

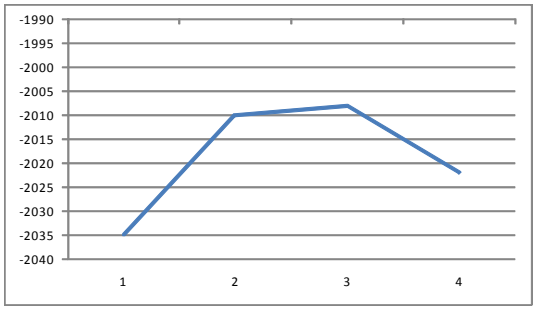

그림 6. 길이(60일)에 따른 BIC 추정
Fig. 6. BIC Estimation by data length(60days)

그림 7과 그림 8을 보면 데이터 길이에 따른 Cheeseman-Stutz(CS) 추정의 결과를 보여준다. 결과를 보면 데이터 객체의 길이가 30일인 경우에는 군집 수를 2개로 잘못된 추정을 보여주고 있으며, 데이터 객체의 길이가 60일인 경우에는 군집 수 3개를 정확히 추정하는 것을 확인할 수 있다.

표 2를 보면 세 종목의 경우에서도 베이지안정보기준(BIC) 추정에서는 데이터 객체 수에 관계없이 군집 수를 3개(3종목) 정확히 추정하는 것을 확인 할 수 있다. 하지만 Cheeseman-Stutz(CS) 추정에서는 데이터 객체가 7개인 경우에는 정확히 군집 수를 추정하나 3개와 5개인 경우에는 군집 수(2종목) 부정확하게 추정하는 것을 확인할 수 있다.

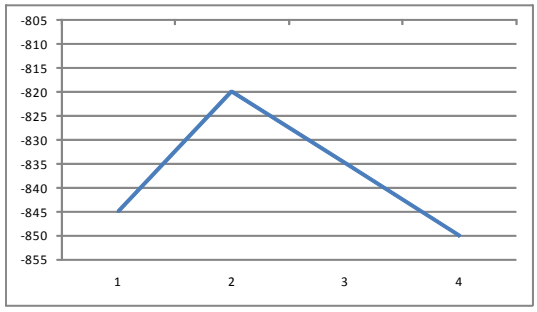

그림 7. 길이(30일)에 따른 CS 추정
Fig. 7. CS Estimation by data length(30days)

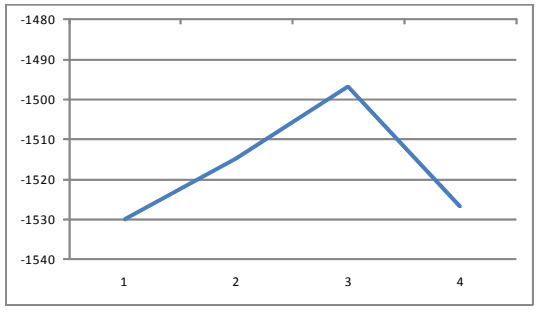

그림 8. 길이(60일)에 따른 CS 추정
Fig. 8. CS Estimation by data length(60days)

표 2. 데이터 객체 수에 따른 추정결과
Table 2. Estimation result by data object number

베이지안정보기준 (BIC) 추정		Cheeseman-Stutz (CS)추정	
3 개체	3 군집	3 개체	2 군집
5 개체	3 군집	5 개체	2 군집
7 개체	3 군집	7 개체	3 군집

위의 두 종목, 세 종목의 경우로 나누어 실험한 결과 대용량의 데이터를 효과적으로 설명하기 위한 군집 수를 추정하기 위한 기준들에 있어 길이와 객체의 수에 따라 베이지안정보기준(BIC) 추정 결과가 Cheeseman-Stutz(CS) 추정 결과보다 정확히 추정하는 것을 확인하였다.

V. 결 론

본 연구에서는 temporal 데이터를 효과적으로 설명할 수 있는 군집 수를 추정하기 위한 휴리스틱 방법론을 살펴보았다.

효율적 추정을 위한 방법론으로 베이지안정보기준(Bayesian Information Criterion)과 Cheeseman-Stutz (CS) 근사법을 살펴보았다.

실제의 주식데이터를 이용하여 실험을 하였으며, 실험결과 Cheeseman-Stutz(CS) 추정의 결과보다 베이지안정보기준(BIC) 추정의 결과가 좀 더 정확히 군집 수를 추정하는 것을 확인하였다.

본 연구에서 제시된 기준을 통하여 기업의 각 업무과정에서 발생되는 다양한 temporal 데이터들에 적용을 시킨다면, 생산 공정에서의 에러의 최소화, 기업의 판매실적의 변화 패턴이나, 고객의 구매행동분석, 주가의 예측 등 많은 영역에서 효과적인 미래의 의사결정을 위한 과정에 적용이 가능할 것이며 좀 더 일반화된 데이터에 적용하기 위한 연구를 통하여 일반적인 분석모델을 세울 수 있는 부분으로 확대가 필요할 것이다.

참 고 문 헌

[1] 오용생, 남도원 ,장지숙, 이동하, 이전영.,“시계열데이터로부터 경향성을 이용한 순차패턴의 탐색”,한국지능정보시스템학회 학술대회 논문집, pp325-332, 2000.

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

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

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

[5] 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.

[6] L. Rabiner, " A tutorial on Hidden Markov Models and selected applications in speech recognition," Proc. of IEEE77, pp.257-286, 1989.

[7] 조영희.,“시계열데이터의 의미기반 패턴매칭과 예측에 관한 연구”, 단국대학교 박사학위논문, 2009.

[8] 전진호.,“시계열데이터의 모델기반 클러스터링을 통한 예측모델 결정에 관한 연구”, 단국대학교 박사학위 논문, 2007.

[9] Cheeseman, P., and Stutz, J. "Bayesian classification(autoclass)" Kluwer Academic Publishers, Vol 70. pp117-126, 1996.

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

저자 소개

전 진 호 (정회원)

• 학위
1994. 관동대학교 경영학과 경영학사
1998. 명지대학교 경영정보학과 경영학석사
2007. 단국대학교 컴퓨터과학 이학박사

• 경력
2009.9 - 현재 관동대학교 경영학과 조교수
<주관심분야 : 데이터마이닝, 기계학습>

김 민 수 (정회원)

• 학위
1997. 관동대학교 무역학과 경영학사
1999. 명지대학교 무역학과 경영학 석사
2004. 명지대학교 무역학과 경영학 박사

• 경력
2009.9 - 현재 관동대학교 호텔경영학과 조교수
<주관심분야 : RFID, 디지털콘텐츠>