시계열 데이타 클러스터링에서 푸리에 진폭 기반의 프라이버시 보호

(Privacy-Preserving Clustering on Time-Series Data Using Fourier Magnitudes)

김 혜 숙 † 문 양 세 ††
(Hea-Suk Kim) (Yang-Sae Moon)

요 약 본 논문에서는 시계열 데이타 클러스터링에서 DFT 진폭 기반의 프라이버시 보호 기법을 제안한다. 기존의 프라이버시 보호 연구인 **DFT 계수 기법**은 원본과 유사한 데이타가 복원될 수 있어 프라이버시 보호 측면에서 큰 문제점이 있다. 반면에, 제안한 DFT 진폭 기법은 DFT 변환 후에 위상을 제외한 진폭만을 사용함으로써 원본 데이타를 복원하기 매우 어려운 특징을 가진다. 본 논문에서는 우선 기존의 **DFT 계수 기법**이 복원이 용이한 함수이고, 제안한 DFT 진폭 기법이 복원이 어려운 함수임을 체계적으로 설명한다. 다음으로, 클러스터링 정확도를 대신하고 진폭을 선택하기 위한 척도로서 **거리-순서** 보존 정도의 개념을 제안한다. 거리-순서 보존 정도는 객체들의 상대적 순서가 클러스터링 보호 함수의 적용 전후에 얼마나 보존되는지의 척도를 나타낸다. 본 논문에서는 이러한 거리-순서 보존 정도의 개념을 사용하여 DFT 진폭 기법에서 진폭을 선택하는 탐욕적 전략들을 제시한다. 즉, 제안한 탐욕적 전략은 거리-순서 보존 정도를 극대화하는 방향으로 DFT 진폭을 선택하여, 궁극적으로 클러스터링 정확도를 높이고자 하는 방법이다. 마지막으로 실험을 통해 제안한 거리-순서 보존 정도가 클러스터링 정확도를 대신할 수 있는 척도임을 보인다. 또한, 제안한 DFT 진폭 기법의 탐욕적 전략들이 기존의 DFT 계수 기법에 비해 정확도가 크게 떨어지지 않음을 확인한다. 이 같은 결과를 볼 때, 제안한 DFT 진폭 기법은 DFT 계수 기법에 비해 프라이버시 보호 정도를 크게 개선했을 뿐 아니라 비교적 정확한 클러스터링 정확도를 보이는 우수한 연구 결과라 사료된다.

키워드 : 시계열 데이타, 클러스터링, 프라이버시 보호, DFT, 푸리에 진폭

Abstract In this paper we propose Fourier magnitudes based privacy preserving clustering on time-series data. The previous privacy-preserving method, called *DFT coefficient method*, has a critical problem in privacy-preservation itself since the original time-series data may be reconstructed from privacy-preserved data. In contrast, the proposed *DFT magnitude method* has an excellent characteristic that reconstructing the original data is almost impossible since it uses only DFT magnitudes except DFT phases. In this paper, we first explain why the reconstruction is easy in the DFT coefficient method, and why it is difficult in the DFT magnitude method. We then propose a notion of *distance-order preservation* which can be used both in estimating clustering accuracy and in selecting DFT magnitudes. Degree of distance-order preservation means how many time-series preserve their relative distance orders before and after privacy-preserving. Using this degree of distance-order preservation we present greedy strategies for selecting magnitudes in the DFT magnitude method. That is, those greedy strategies select DFT magnitudes to maximize the degree of distance-order preservation, and eventually we can achieve the relatively high clustering accuracy in the DFT magnitude method. Finally, we empirically show that the degree of distance-order

· 이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2007-331-D00381)

† 학생회원 : 강원대학교 컴퓨터과학과
hskim@kangwon.ac.kr
†† 종신회원 : 강원대학교 컴퓨터과학과 교수
ysmoon@kangwon.ac.kr

논문접수 : 2008년 3월 24일
심사완료 : 2008년 9월 19일

정보과학회논문지: 데이타베이스 제35권 제6호(2008.12)

preservation is an excellent measure that well reflects the clustering accuracy. In addition, experimental results show that our greedy strategies of the DFT magnitude method are comparable with the DFT coefficient method in the clustering accuracy. These results indicate that, compared with the DFT coefficient method, our DFT magnitude method provides the excellent degree of privacy-preservation as well as the comparable clustering accuracy.

Key words : Time-series data, Clustering, Privacy preserving, DFT, Fourier magnitude

1. 서 론

컴퓨터가 처리하는 데이타 양이 증가하고 그 종류가 다양해짐에 따라 개인 데이타의 프라이버시 보호에 대한 문제가 대두되고 있다. 프라이버시 보호는 누가 개인에 관한 정보를 수집 및 관리할 것이며, 정보의 제공이 얼마나 안전한가에 관련한 문제이다. 이러한 프라이버시 보호 문제는 대용량 데이타를 대상으로 하는 데이타 마이닝의 여러 분야에서도 활발한 연구가 이루어져왔다[1-9]. 데이타 마이닝에서의 프라이버시 보호 문제는 정보제공자에 의해 제공된 정보 중 민감한 개인 정보의 노출이 없이도 가능한 정확한 마이닝 결과를 얻는 것이다. 본 논문에서는 이 중 시계열 데이타를 대상으로 하는 프라이버시 보호 클러스터링 문제를 다룬다.

1990년부터 연구되기 시작한 데이타 마이닝에서의 프라이버시 보호 연구[1,2]는 크게 1) 원본 데이타를 왜곡하거나 변환하는 데이타 교란(perturbation)[3,4,7]과 2) 분산 환경에서 사용할 수 있는 안전한 다원 계산(secure multiparty computation: SMC)[2,5,6]으로 구분할 수 있다. 데이타와 관련해서는 Mukherjee와 Chen[8]이 DFT 기반의 클러스터링을 위한 프라이버시 보호 기법을 제안하였다. 그러나, 이 방법은 DFT 역변환을 통해 원본과 유사한 데이타가 복원될 수 있다는 문제점이 있다. 따라서, 본 논문에서는 프라이버시 보호된 데이타가 노출되어도 원래 데이타로의 복원이 어려울 뿐 아니라 마이닝 결과가 비교적 정확한 새로운 프라이버시 기법을 제안한다.

본 논문에서는 먼저 시계열 데이타의 프라이버시 보호 정도와 관련하여 복원이 용이한 함수와 복원이 어려운 함수의 개념을 정의한다. **복원이 용이한** 함수란 프라이버시 보호된 데이타를 기반으로 원본과 유사한 데이타를 복원할 수 있는 함수이다. 반대로, **복원이 어려운** 함수는 프라이버시 보호된 데이타가 노출되어도 원본 데이타를 복원하기 어려운 함수이다. 즉, 복원이 어려운 함수가 프라이버시 보호 측면에서 우수한 함수라 할 수 있다. 또한, 본 논문에서는 클러스터링 정확도를 비교하기 위하여 **거리-순서**(*distance-order*) 보존 정도의 개념을 제시한다. 거리-순서 보존 정도는 원본 데이타에서 객체들의 상대적 순서가 프라이버시 보호된 이후에도 유지되는 정도로써, 이를 사용하여 클러스터링의 정확도를 대신할 수 있다. 결국, 본 논문에서는 거리-순서 보존 정도를 높이면서도 원본 데이타의 복원이 어려운 프라이버시 보호 기법을 제안한다.

시계열 데이타의 클러스터링을 위한 프라이버시 보호 기법으로 본 논문에서는 **DFT 진폭 기법**을 제안한다. 기존 연구[8]에서는 DFT 계수를 사용하였는데, 앞서 설명한 바와 같이 이 방법은 원본과 유사한 데이타가 복원될 수 있는 문제점이 있다. 본 논문에서는 이러한 기존의 **DFT 계수 기법**이 복원이 용이한 문제점이 있음을 체계적으로 설명한다. DFT 계수 기법의 복원 용이성 문제를 해결하기 위하여, 본 논문에서는 DFT 진폭만을 클러스터링에 사용하는 DFT 진폭 기법을 제안한다. 이는 시계열 데이타를 DFT 변환한 후 결과인 위상과 진폭의 함수에서 위상을 제외한 진폭만을 사용하는 개념이다. 본 논문에서는 DFT 진폭 기반의 프라이버시 보호 기법이 복원이 어려운 장점을 가짐을 체계적으로 주장한다. 즉, DFT 결과에서 위상을 제외하고 진폭만을 사용할 경우 원본 시계열의 복원이 매우 어려움을 체계적으로 설명한다. 결과적으로, 제안한 DFT 진폭 기법은 프라이버시 보호 측면에서 우수하다 할 수 있다.

DFT 진폭 기법의 클러스터링 정확도를 높이기 위해 본 논문에서는 DFT 진폭을 선택하는 여러 가지 전략을 제안한다. 시계열 사이의 거리는 DFT 진폭에 많은 영향을 받으나[10], 위상에 의해서도 적잖은 영향을 받게 된다. 따라서, DFT 진폭 기법에 있어서 어떠한 진폭을 선택하여 클러스터링에 사용하느냐가 거리-순서 보존 정도를 높이고, 더 나아가 클러스터링 정확도를 높이는 중요한 역할을 한다. 먼저, 가장 간단한 방법으로, DFT 계수 기법에서와 같이 에너지가 집중된 처음 몇 개의 진폭을 선택하는 **순차 선택 전략**(*sequential selection strategy*)을 제안한다. 그런데, 처음 몇 개의 진폭(혹은 계수)을 선택하는 이러한 전략은 DFT 계수 기법에서는 널리 사용되었으나 DFT 진폭의 경우에도 그대로 적용되는지 확인된 바가 없다. 더구나, 순차 선택 전략은 거리-순서 보존 정도를 전혀 고려하지 않은 문제점이 있다. 이에 따라, 본 논문에서는 거리-순서 보존 정도를 극대화하기 위한 두 가지 탐욕적 전략을 제안한다. 첫째는 **단순 탐욕 전략**(*simple greedy strategy*)으로서, DFT

진폭 각각이 가지는 거리-순서 보존 정도를 샘플 시계열 데이타를 대상으로 계산한 후, 보존 정도가 큰 진폭들을 선택하는 전략이다. 둘째는 **심화 탐욕 전략**(*advanced greedy strategy*)로서, 단순 탐욕 전략으로 하나의 진폭을 선택하고 그 기반 하에서 다음 진폭을 선택하는 루프를 반복하여 원하는 수의 진폭을 선택하는 보다 탐욕적인 전략이다. 이러한 탐욕적 전략은 진폭을 선택하는데 있어서는 많은 시간이 걸리나, 거리-순서 보존 정도를 극대화하여 궁극적으로 클러스터링 정확도를 높이게 된다.

본 논문에서는 다양한 시계열 데이타 집합을 대상으로 한 실험을 통해 제안한 DFT 진폭 기법이 기존의 DFT 계수 기법에 견줄만한 실용적인 방법임을 입증한다. 실험 결과는 크게 세 가지로 요약할 수 있다. 첫째, 제안한 거리 순서 보존 정도는 클러스터링 정확도를 대신할 수 있는 우수한 척도로 확인되었다. 둘째, 거리-순서 보존 정도 측면에서 제안한 DFT 진폭 기법은 DFT 계수 기법에 비해 크게 뒤지지 않았으며(평균 87% 보존), 이에 따라 클러스터링 정확도에 있어서도 큰 차이를 보이지 않았다(평균 79% 보존). 셋째, 제안한 세 가지 전략에 있어서 탐욕적 전략이 순차 선택 전략에 비해 우수한 결과를 보였고, 심화 탐욕 전략이 단순 탐욕 전략보다 우수한 결과를 보였다. 이 같은 결과를 종합할 때, 제안한 DFT 진폭 기법은 프라이버시 보호 특성이 매우 우수함과 동시에, 마이닝 결과의 정확도 측면에서도 DFT 계수 기법에 비해 크게 뒤지지 않는 우수한 프라이버시 보호 기법이라 할 수 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 클러스터링과 데이타 마이닝에서의 프라이버시 기법에 관련된 연구를 소개한다. 제3장에서는 시계열 데이타 클러스터링에서의 프라이버시 보호 문제를 정의한다. 제4장에서는 프라이버시 보호 정도와 거리-순서 개념을 정의하고, DFT 진폭 기반의 프라이버시 보호 기법을 제안한다. 제5장에서는 클러스터링 결과의 정확도를 높이기 위한 DFT 진폭 선택 전략을 제안한다. 제6장에서는 실험 결과를 설명하고, 마지막으로 제7장에서 결론을 맺는다.

2. 관련 연구

클러스터링은 데이타 마이닝, 통계, 기계학습 등 여러 학문분야에서 매우 다양한 방법들이 연구되었다. 클러스터링은 두 객체간의 유사성(similarity)을 계산하여 객체들을 유사한 특성을 가진 클러스터로 구분하는 기법이다. 이러한 클러스터링 알고리즘들은 분할 알고리즘(partitioning algorithm)과 계층 알고리즘(hierarchical algorithm)으로 구분할 수 있다[11]. 분할 알고리즘에서는 클러스터링을 수행할 객체의 수와 생성할 클러스터의 수가 미리 주어진다. 그러면, 클러스터의 수에 따라 각 클러스터의 중심값을 결정하고, 중심값을 제외한 나머지 객체들을 각 클러스터로 분할하게 된다. 이때 객체의 분할 기준은 해당 객체와 그 객체가 속하게 될 클러스터의 중심값과의 유사성이다. 분할 알고리즘은 이런 과정을 반복하면서 최종 클러스터를 형성하게 된다. 대표적인 분할 알고리즘으로는 k-means[12], PAM[13], CLARA[13], CLARANS[14] 등이 있다. 다음으로, 계층 알고리즘은 각 객체(혹은 객체들의 그룹)들을 계층적으로 분해(divisive)하거나 통합(agglomrative)하는 과정을 사용자가 원하는 수준에 이를 때까지 반복하면서 클러스터링을 수행한다. 대표적인 계층 알고리즘으로는 BIRCH[15], CURE[16] 등이 있다. 기존의 이러한 클러스터링 연구와 본 논문의 연구는 직교적(orthogonal)이다. 즉, 기존 연구들은 본 논문에서 다루는 프라이버시 보호 기법에 활용할 수 있으며, 본 논문에서는 이 중 가장 대표적인 k-means 알고리즘을 클러스터링 기법으로 이용한다.

Agrawal과 Srikant[1], Lindell과 Pinkas[2]에 의해 데이타 마이닝의 프라이버시 보호 문제가 처음으로 제안된 후, 데이타 마이닝 분야에서도 프라이버시 보호 연구가 많이 시도되고 있다. 최근까지 진행된 데이타 마이닝에서의 프라이버시 보호 연구는 크게 두 분야로 구분할 수 있다. 첫 번째는 데이타 교란을 사용한 프라이버시 보호 기법이다[3,4,7]. 이 기법은 각 객체의 정보를 보호하기 위해 원본 데이타에 임의의 값 추가(randomization)[4], 원본 데이타의 왜곡(distortion)[3], 원본 데이타의 변환(transformation)[7] 등의 전처리 과정을 수행한 후 마이닝을 수행하는 기법이다. 이러한 데이타 교란 기법은 비교적 간단히 사용할 수 있는 장점이 있으나, 원본 데이타가 하나라도 노출되어 회귀분석(regression)이나 웨이블릿 기반(wavelet-based) 필터링의 과정을 거치면 원본과 유사한 데이타가 복원될 수 있다는 단점이 있다[9]. 또한, 이들 방법을 유클리디안 거리 기반의 클러스터링이나 분류(classification)에 사용할 경우 객체간 거리를 보장하지 못하는 단점이 있어[8], 본 논문에서 다루는 시계열에는 적용하기 어렵다. 두 번째는 SMC 기법을 사용한 프라이버시 보호 기법이다[2,5,6]. 이 SMC 기법은 마이닝을 수행할 데이타들이 분산되어 있을 때 사용하는 프라이버시 보호 기법이다. 먼저, 분산되어 있는 각 부분별로 마이닝을 수행한 후, 그 결과를 다른 부분과 공유하거나 최종 사이트로 전송하며, 이 과정에 암호화 기법이 사용될 수 있다. 다음으로, 최종 사이트는 각 부분에서 전송된 중간 결과를 집계하여 최종 마이닝 결과를 도출한다. 이러한 SMC 기법은 데이타 전송 시 악의적인 사용자에 의해 데이타가 노출될 수 있다는 단

점이 있다. 이 방법은 분산 환경에 초점을 맞춘 프라이버시 보호 기법으로, 본 논문에서 수행하는 시계열 대상의 프라이버시 보호 문제와는 직교적이라 할 수 있다.

Mukherjee와 Chen[8]은 시계열 데이타 대상의 클러스터링(혹은 분류)을 위한 프라이버시 보호 기법을 최근 제안하였다. 일반적으로 시계열 데이타 대상의 클러스터링에서는 두 시계열간의 유클리디안 거리를 유사성 척도로 사용하였다[17]. 그런데, 데이타 교란을 사용하는 기존의 프라이버시 보호 기법[4,7,3]은 시계열 간의 유클리디안 거리를 보존하지 못하므로, k-means나 k-NN 등의 기존 클러스터링 알고리즘에 그대로 사용할 수 없다. 이 점에 착안하여 Mukherjee와 Chen은 DFT(혹은 DCT)를 사용한 시계열 대상의 프라이버시 보호 기법을 제안하였다. 이들은 원본 시계열을 DFT로 변환한 후, 변환된 시계열에서 에너지가 집중된 몇 개의 계수만을 클러스터링에 사용하는 방법을 제안하였다. 즉, 원본 시계열 자체를 사용하는 것이 아니라 이의 특성을 반영하는 몇 개의 계수만을 거리 계산에 사용함으로써, 클러스터링 과정에 있어서 원본 시계열을 보호하는 것이다. 이러한 계수 추출 방법은 시계열 매칭 분야[18,19,20]에서 저차원 변환으로도 잘 알려져 있으며, 추출된 계수가 원래 시계열의 거리 순서를 어느 정도 보존한다는 특징을 갖는다.

그런데, Mukherjee와 Chen의 이러한 DFT 기반 방법은 프라이버시 보호 측면에서 큰 문제점이 있다. 이는 DFT의 경우, 역변환을 통해서 원본과 유사한 시계열이 복원될 수 있기 때문이다. 클러스터링에 사용되는 DFT 계수는 원본 시계열의 에너지가 집중된 것으로서, 이들 몇 개의 계수만을 사용해도 원본과 유사한 시계열이 복원될 수 있는 것이다. 이러한 문제점은 다음 제3장에서 구체적으로 설명하며, 본 논문에서는 이와 같이 계수가 노출되어도 원본 시계열로의 복원이 어려운 강력한 프라이버시 기법을 제안한다.

3. 문제 정의

본 연구의 동기는 기존의 DFT 계수 기법[8]이 프라이버시 보호에 문제가 있다는 직관에서 출발한다. 즉, DFT 계수 기법의 경우, DFT 계수와 이들의 위치가 알려진다면 원본 시계열이 유사하게 복원될 수 있다는 문제점을 가지고 있다. 그림 1은 이러한 예로서, 길이가 271인 원본 시퀀스와 이를 DFT 변환한 후 높은 에너지를 갖는 계수를 각각 두 개, 세 개, 네 개를 선정한 다음, DFT 역함수를 사용하여 복원한 시퀀스를 나타낸다. 그림에서 보면, 계수 세 개를 사용한 경우, 원본 시퀀스와 거의 유사한 추세를 가지는 시퀀스가 복원되었다. 더구나, 계수 네 개를 사용한 경우에는 원본 시퀀스

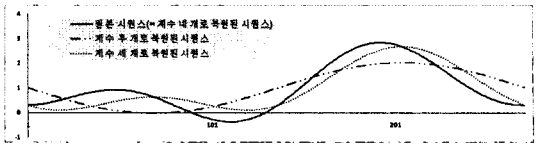


그림 1 DFT 계수 방법에서의 원본 시퀀스와 복원된 시퀀스

가 100% 정확하게 복원되었다. 그림 1의 예에서 보듯이, 단지 몇 개의 DFT 계수만을 사용하더라도 원본 시퀀스와 매우 유사한 시퀀스가 복원됨을 알 수 있다. 이와 같이, DFT 계수를 직접 사용하는 기존 기법은 프라이버시 보호에 심각한 문제가 있음을 알 수 있다.

데이타 마이닝 프라이버시 보호 연구들의 공통적인 목표는 크게 두 가지이다. 첫째, 프라이버시 보호 기법이 적용된 데이타로 인한 복원이 어려워야 한다는 것이다. 즉, 프라이버시 보호된 데이타가 악의적인 사용자에게 노출이 될 수 있는 데, 이때 원본 데이타의 복원이 어려워야 한다는 것이다. 둘째, 프라이버시 보호된 데이타를 대상으로 수행한 마이닝 결과의 정확성이 높아야 한다는 것이다. 즉, 프라이버시 보호 기법을 적용한 데이타를 마이닝한 결과는 프라이버시 보호 기법을 적용하지 않은 원본 데이타를 마이닝한 결과와 유사해야 한다는 것이다. 그런데, 이 두 가지 목표는 시계열 데이타 대상의 프라이버시 보호 클러스터링에서도 쉽게 달성할 수 없는 어려운 문제이다. 앞서도 설명하였듯이, 시계열 데이타 마이닝의 저차원 변환에 주로 사용되는 DFT나 PAA 등은 원본 데이타와 유사하게 복원이 가능하기 때문에 프라이버시 보호 기법으로 사용하기 어렵다. 게다가 원본 데이타의 프라이버시 보호 정도를 높이기 위해 보다 저차원으로 변환할수록 마이닝 결과의 정확도가 떨어지게 되는 문제점이 있다. 결국, 시계열 데이타의 복원 정도와 클러스터링 결과의 정확도는 상충(trade-off) 관계에 있다. 이에 따라, 본 연구에서는 시계열 데이타 대상의 프라이버시 보호 클러스터링 문제를 다음과 같이 개념적으로 정의한다.

문제 정의: 시계열 데이타 대상의 프라이버시 보호 클러스터링 기법은 다음 두 가지 특성을 가져야 한다. 첫째, 프라이버시 보호된 데이타를 사용하여 원본과 유사한 데이타를 복원하기 어려워야(혹은 불가능해야) 한다. 둘째, 프라이버시 보호된 데이타를 사용한 클러스터링 결과가 원본 데이타를 사용한 클러스터링 결과와 유사해야 한다.

4. DFT 진폭 기반의 프라이버시 보호 클러스터링

본 장에서는 DFT 계수 대신 DFT 진폭을 사용하는 프라이버시 보호 클러스터링 기법을 제안한다. 제4.1절

에서는 프라이버시 보호 정도를 정의하고, 프라이버시 보호에 따른 클러스터링 정확도를 설명한다. 제4.2절에서는 본 논문에서 프라이버시 보호를 위해 사용하는 DFT 진폭 기법을 자세히 설명한다. 제4.3절에서는 제4.2절에서 제안한 DFT 진폭 기법을 복원 성질의 관점에서 설명한다.

4.1 프라이버시 보호 정도와 클러스터링 정확도

기존 데이타 마이닝 프라이버시 보호 기법들과 같이 본 연구의 첫 번째 목표는 프라이버시 보호된 데이타로부터 원본 데이타의 복원을 어렵게 하는 것이다. 본 논문에서는 이러한 프라이버시 보호 정도를 변환된 데이타로부터 원본 시계열의 복원 정도로 평가하고자 한다. 그림 2는 이러한 프라이버시 보호 정도를 개념적으로 나타낸 것이다. 그림을 보면, 원본 시계열 S는 주어진 프라이버시 보호 함수 P에 의해 프라이버시 보호된 데이타 S_p로 변환된다. 그런 다음, 이렇게 프라이버시 보호된 데이타 S_p를 사용하여 데이타 마이닝을 수행한다. 그런데, 이때 프라이버시 보호된 데이타 S_p를 원본과 유사한 시계열로 복원하는 함수 P^{-1}이 존재한다고 하자. 그리고, P^{-1}에 의해 복원된 시계열 S^{-1}이 원본 시계열 S와 유사할 경우 P는 프라이버시 보호가 잘 되지 않는 함수이며, 그렇지 않은 경우 P는 프라이버시 보호가 잘 되는 함수라 할 수 있다.

본 논문에서는 이와 같이 프라이버시 보호가 잘되는 함수와 그렇지 않은 함수를 구분하기 위하여, 원본 시계열의 복원 용이성을 다음과 같이 정형적으로 정의한다.

정의 1: 길이가 n인 시계열을 S = {$S[0]$, …, $S[n-1]$}이라 하고, 프라이버시 보호 함수 P를 적용하여 얻은 길이 $c(\leq n)$인 시계열을 S_p = {$S_p[1]$, …, $S_p[c]$}라 하며, P의 복원 함수 P^{-1}가 있어 S_p로부터 시계열 S^{-1}= {$S^{-1}[0]$, …, $S^{-1}[n^{-1}]$}을 복원한다고 하자. 이때, P^{-1}에 의해 생성된 S^{-1}이 원본 시계열 S와 유사하다면, P는 **복원이 용이한** 함수라 정의한다. 반면에, S^{-1}이 S와 유사하지 않거나, 그러한 복원 함수 P^{-1}가 존재하지 않는다면, P는 **복원이 어려운** 함수라 정의한다.

상기 정의에서, 복원의 용이함이나 어려움을 주장하기 위한 S와 S^{-1}의 유사한 정도는 주관적인 요소이다. 따라서, 본 논문에서는 이를 직관적이고 일반적인 유사성으로 대신하기로 하며, 이에 대한 자세한 논의는 생략한다. 제3장에서 언급한 바와 같이 DFT 계수 기법[8]은 프라이버시 보호된 데이타로부터 원본과 유사한 시계열이 손쉽게 복원되므로, 다음과 같이 주장(assertion)할 수 있다.

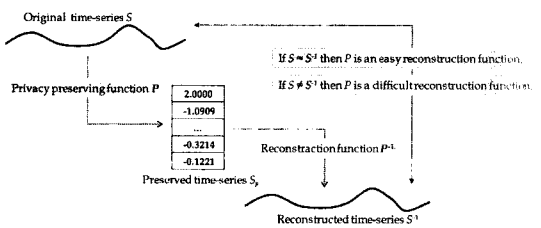


그림 2 프라이버시 보호 함수의 복원이 용이한 정도와 복원이 어려운 정도

주장 1: 시계열을 DFT로 변환한 후 몇 개의 계수를 취하는 프라이버시 보호 함수는 복원이 용이한 함수이다.

증명: 제3장에서 설명한 바와 같이 에너지가 집중된 몇 개의 DFT 계수로 원본과 매우 유사한 시계열을 복원할 수 있다. 따라서, DFT 계수를 사용하는 프라이버시 보호 함수는 복원이 용이한 함수라 주장할 수 있다.

주장 1에 따라 DFT 계수를 사용하는 기존 연구[8]의 프라이버시 보호 함수는 복원이 용이한 함수라 할 수 있고, 결국 프라이버시 보호 측면에서는 좋은 함수라 할 수 없다.

프라이버시 보호 마이닝에서 다음으로 고려해야 하는 사항은 프라이버시 보호된 데이타로부터 얻은 마이닝 결과가 얼마나 정확하냐는 것이다. 즉, 우리 문제의 경우 원본 시계열 데이타에서 얻은 클러스터링 결과와 프라이버시 보호된 데이타에서 얻은 클러스터링 결과에 대한 비교가 필요하다. 이를 위한 가장 간단한 방법은 원본 및 프라이버시 보호된 데이타에 대해 각각 클러스터링을 수행한 후 그 결과를 직접 비교하는 것이다. 그러나 이 방법을 사용하기 위해서는, 시간이 많이 소모되는 클러스터링 작업을 여러 프라이버시 보호 기법 및 여러 데이타 집합 각각에 대해 매번 수행해야 하는 어려움이 따른다. 이에 따라, 본 논문에서는 클러스터링 결과를 직접 비교하는 대신, 다음과 같이 시계열간 유사성의 상대적 순서를 나타내는 거리-순서의 개념을 제시하고, 이를 프라이버시 보호된 데이타로부터 얻은 클러스터링 결과의 정확도 대신 사용한다.

정의 2: 시계열 Q, $S1$, $S2$가 주어졌고, 이들 간 거리를 $D(Q,S1)$과 $D(Q,S2)$라 하자. 또한, 이들 시계열에 프라이버시 보호 함수 P를 적용한 결과 시퀀스를 Q_p, $S1_p$, $S2_p$라 하고, 이들 간 거리를 $D(Q_p,S1_p)$과 $D(Q_p,S2_p)$라 하자. 이때, 거리 $D(Q,S1)$과 $D(Q,S2)$의 상대적 순서가 변환 후 거리인 $D(Q_p,S1_p)$과 $D(Q_p,S2_p)$의 상대적 순서와 같다면, P는 **거리-순서**(*distance-order*)를 보존한다고 정의한다. 즉, 다음 식 (1) 또는 (2)가 성립하면, P는 거리-순서를 보존한다고 정의한다.

$$D(Q,S1) \leq D(Q,S2) \Rightarrow D(Q_p,S1_p) \leq D(Q_p,S2_p) \quad (1)$$

$$D(Q,S1) \geq D(Q,S2) \Rightarrow D(Q_p,S1_p) \geq D(Q_p,S2_p) \quad (2)$$

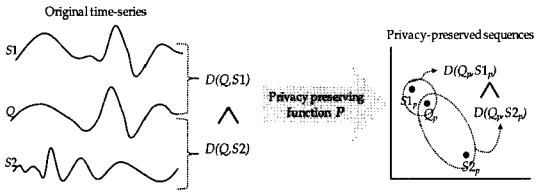


그림 3 거리-순서를 보존하는 함수의 예

그림 3은 정의 2의 거리-순서 보존의 개념을 직관적으로 나타낸 것이다. 그림을 보면, 원본 시계열 Q와 $S1$의 거리는 Q와 $S2$의 거리보다 가까우며, 프라이버시 보호 함수 P에 의해 변환된 시퀀스 Q_p와 $S1_p$의 거리 역시 Q_p와 $S2_p$의 거리보다 가깝다. 이와 같이 함수 P의 적용 전후의 상대적 순서가 변화하지 않으므로, 함수 P는 거리-순서를 보존한다고 할 수 있다. 거리-순서의 개념을 도입한 이유는 대부분의 클러스터링 기법이 시계열 간 상대적 거리인 거리-순서에 기반하여 클러스터를 식별하기 때문이다. 이와 같은 거리-순서를 사용했을 때의 장점은 1) 거리-순서는 특정 클러스터링 기법에 의존적이지 않는 객관적인 척도이고, 2) 거리-순서의 보존 여부는 계산을 통해 쉽게 판단된다는 점이다.

4.2 DFT 진폭의 개념

본 논문에서는 시계열 데이타의 프라이버시 보호를 위하여 DFT 진폭(magnitude)을 사용한다. 하나의 시계열은 DFT를 통해 여러 주파수(frequency)들의 합으로 표현되는데, 이때 각 주파수는 진폭과 위상(phase)의 함수로 나타난다. 이중 위상을 제외하고 진폭만을 사용하는 개념은 윤곽선 이미지 매칭[21]에서 회전 불변을 지원하기 위해서도 사용된바 있는데[10], 본 논문에서는 이를 시계열의 프라이버시 보호에 사용하고자 한다. 길이 n인 시계열 $X(=\{X[0], \cdots, X[n-1]\})$에 대한 DFT 진폭 시퀀스 X_{DFTm}은 다음의 식 (3)과 식 (4)를 통해 구할 수 있다. 다음 식 (4)에서 $\|X_{DFT}[i]\|$은 복소수 $X_{DFT}[i]$의 놈(norm)을 나타낸다.

$$X_{DFT}[k]=\frac{1}{n}\sum_{i=0}^{n-1}X[i]e^{\frac{-2\pi ik}{n}\cdot j}, \quad (3)$$

$$where\ k=0, 1, ..., n-1;$$

$$X_{DFTm}[i]=\|X_{DFT}[i]\|,\ i=0, 1, ..., n-1 \quad (4)$$

그림 4는 프라이버시 보호[8] 및 저차원 변환[18,19, 22] 등의 기존 연구에서 사용된 DFT 계수와 본 논문에서 사용하는 DFT 진폭과의 관계를 나타낸 것이다. 기존 연구에서는 그림에 나타난 DFT 계수인 a와 b를 직접 사용하므로, 원본 시계열의 복원이 용이하다. 반면에, 본 논문에서는 a와 b의 놈인 진폭 m만을 사용하므로, 원본 시계열의 복원이 어렵게 된다. 즉, 원본 시계열을

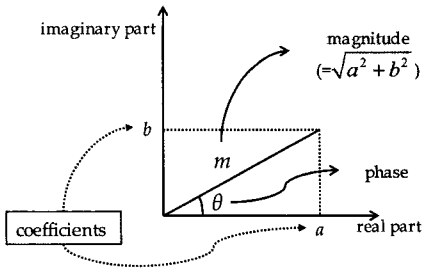


그림 4 DFT 진폭

복원할 주요 정보인 위상(그림에서 θ에 해당)을 사용하지 않기 때문에 원본 시계열의 복원이 어렵게 되는 것이다.

그런데, DFT 진폭을 프라이버시 보호 함수로 사용하기 위해서는 1) DFT 진폭으로 인한 원본 시계열의 복원이 어려움을 확인해야 하고, 2) DFT 진폭만을 사용했을 경우의 클러스터링 결과가 어느 정도 정확한지 확인해야 한다. 다음의 제4.3절에서는 DFT 진폭이 정의 1에서 제시한 복원의 용이성을 가지지 않음을, 즉 복원이 어려움을 확인한다. 또한, 제6장의 실험에서는 정의 2의 거리-순서의 개념을 사용하여 DFT 진폭이 DFT 계수에 비교할만한 클러스터링 정확도를 제공함을 확인한다.

4.3 DFT 진폭의 복원 성질

DFT 계수와 달리 DFT 진폭만을 사용하는 프라이버시 보호 함수가 복원이 어려운 함수임을 다음과 같이 주장한다.

주장 2: 시계열을 DFT로 변환한 후 몇 개의 진폭을 취하는 프라이버시 보호 함수는 복원이 어려운 함수이다.

증명: 그림 4에 따르면, DFT 계수 a와 b는 진폭 m과 위상 θ와 동일한 정보를 가진다. 즉, a와 b가 주어지면 m과 θ를 구할 수 있고, 반대로 m과 θ가 주어지면 a와 b를 구할 수 있다. 주장 1에서 설명한 바와 같이 계수 a와 b를 사용하는 DFT 계수 기반 함수는 복원이 용이한 함수였다. 따라서, DFT 진폭 기반 함수가 복원이 용이한 함수가 되려면, 주어진 진폭만으로 a와 b를 복원할 수 있어야 한다. 즉, 진폭 m으로 원본과 유사한 시계열을 복원할 수 있어야 한다. 그러나, DFT 진폭 기반 함수의 경우, 주어진 진폭 m만으로 구할 수 있는 계수 a와 b는 무수히 많다. 즉, 위상이 주어지지 않은 상태에서 진폭만 가지고는 무수히 많은 a와 b가 존재할 수 있고, 결국 무수히 많은 시계열이 생성된다. 이는 진폭이 주어진 반면에 위상이 주어지지 않아, 모든 가능한 위상 값(0°~360°)을 고려해야 하기 때문이다. 결국, DFT 진폭 기반 함수에서 주어진 진폭만으로는 원본 시계열과 유사한 시계열을 복원하기 어렵게 된다. 이에 따라

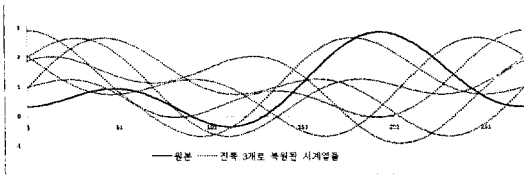


그림 5 원본 시퀀스와 DFT 진폭으로 복원된 시퀀스

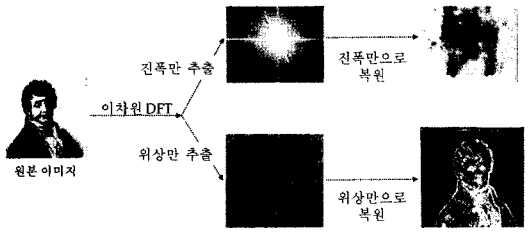


그림 6 이미지 처리에서의 DFT 역변환

DFT 진폭 기반 함수는 복원이 용이하지 않은, 즉 복원이 어려운 함수라 주장할 수 있다.

주장 2에 따라 DFT 진폭을 사용하는 본 논문의 프라이버시 보호 함수는 복원이 어려운 함수라 할 수 있고, 결국 프라이버시 보호 측면에서는 좋은 함수라 할 수 있다.

그림 5는 그림 1의 예에서, DFT 계수 대신에 DFT 진폭만을 사용했을 경우에 원본 시계열에서 복원된 시계열을 나타낸다. 그림에서 보듯이, 위상 값에 따라 수많은 시계열로 복원될 수 있는데, 실제로 위상 값을 알 수 없기 때문에, 이 중 어떤 시계열로 복원할지를 결정할 수 없게 된다. 결과적으로, DFT 진폭만을 사용해서는 원본과 유사한 시계열을 복원하기 어렵고, 이에 따라 주장 2에서와 기술한 바와 같이, DFT 진폭 기반의 프라이버시 보호 함수는 복원이 어려운 함수라 할 수 있다.

DFT 진폭만으로 원본 시계열을 복원하기 어려움은 이미지 데이타의 예에서 개념적으로 확실하게 확인할 수 있다. 그림 6은 이미지 처리 분야에서 이차원 DFT를 수행한 후, 하나는 진폭만을, 다른 하나는 위상만을 사용하여 원본 이미지의 복원을 수행한 예이다[23]. 그림에서 보듯이, 위상만을 사용한 경우는 원본과 유사한 이미지가 복원되는 반면에, 진폭만을 사용한 경우는 원본과 유사한 이미지가 복원되지 않음을 알 수 있다. 이와 유사한 개념으로 일차원 데이타인 시계열에서도 진폭만으로는 원본과 유사한 이미지를 복원하기 어렵게 된다.

5. DFT 진폭 선택 전략

DFT 진폭을 프라이버시 보호에 사용하기 위해서는 어떤 진폭을 선택할지에 대한 전략이 필요하다. 즉, 길이 n인 시계열에서 총 n개의 진폭이 나오는데, 이들 중

Algorithm *sequential-selection* (a time-series S of length n, the number f of DFT magnitudes)
(1) Extract n DFT magnitudes from S by DFT;
(2) Select the first f magnitudes from the DFT magnitudes;

그림 7 순차 선택 전략 알고리즘

f개를 어떤 기준에 의해 선택할 지 전략을 세워야 한다. 그리고 이러한 DFT 진폭 선택 전략은 제3장에서 제시한 문제 정의에 있어서 클러스터링 정확도를 높이기 위한 방향으로 진행하게 된다. 또한 이는 프라이버시 보호 마이닝의 두 번째 목표인 프라이버시 보호에 의한 마이닝 결과의 왜곡을 최소화하려는 노력으로도 해석할 수 있다. 본 장에서는 간단한 선택 전략과 함께 클러스터링 정확도를 높이기 위한 탐욕적 전략들을 제안한다.

가장 쉽게 생각할 수 있는 선택 전략은 처음의 진폭 f개를 선택하는 것이다. 이는 시계열의 많은 에너지가 DFT 계수 처음 몇 개에 집중된다는 성질[18,19]에 따른 것으로, 가장 간단하면서도 직관적으로 우수한 정확도를 나타낼 수 있는 전략이다. 본 논문에서는 이를 **순차 선택 전략**(*sequential selection strategy*)이라 부른다. 즉, 순차 선택 전략은 길이 n인 시계열에서 DFT 진폭 n개를 추출한 후, 처음 f개의 진폭을 순서에 따라 선택한다. 그림 7은 이러한 순차 선택 전략을 수행하는 알고리즘을 나타낸다. 알고리즘에서는 길이 n인 시계열에서 총 n개의 진폭을 추출하고(라인 (1)), 이중 처음 f개의 진폭만을 선택하는(라인 (2)) 매우 간략한 구조를 가진다.

순차 선택 전략은 진폭을 선택하는데 있어서 클러스터링 정확도를 고려하지 않은 문제점이 있다. DFT에 있어서 많은 에너지가 집중되는 처음 몇 개의 계수를 사용하는 것이 효과적이라는 것은 이미 알려진 사실이다. 그러나, 이러한 성질이 DFT 계수가 아닌 진폭을 사용한 경우, 즉 진폭을 사용하여 클러스터링을 수행하는 경우에도 높은 정확도를 보이는지는 확인된 바가 없다. 이에 따라, 본 논문에서는 DFT 진폭을 선택하는데 있어서, 순차 선택 전략의 비교 대상으로 **단순 탐욕 전략**(*simple greedy strategy*)과 **심화 탐욕 전략**(*advanced greedy strategy*)의 두 가지 탐욕적 전략을 제안한다. 제안하는 탐욕적 전략은 제4.1절에서 정의한 거리-순서의 보존 정도를 극대화하는 방향으로 DFT 진폭을 선택한다. 이를 위해 본 논문에서는 각 선택 전략에 대한 **거리-순서 보존 정도**(*degree of distance-order preserving*)를 다음 과정으로 구한다. 먼저, 시계열 데이타베이스에서 임의의 시계열 네 개를 포함하는 샘플들을 구성한다. 그런 다음, 해당 선택 전략이 정의 2의 거리-순서를 보존하는지 확인하는 실험을 모든 샘플을 대상으로 반복한다. 이 때, 거리-순서 보존 정도는 실험을 수행한 전체 샘플 수에 대한 거리-순서가 보존된 샘플 수로 계산

되며, 본 논문에서는 이를 *ddop*라 간략히 표기한다. 즉, 다음 식 (5)와 같이 *ddop*를 계산한다.

$$ddop = \frac{\text{the number of distance} - \text{order preserved samples}}{\text{the total number of samples}} \quad (5)$$

먼저, 단순 탐욕 전략은 DFT 진폭 각각에 대해서 거리-순서 보존 정도를 구한 후, 보존 정도가 큰 처음 *f*개를 선택하는 전략이다. 단순 탐욕 전략의 DFT 진폭 선택 과정을 자세히 설명하면 다음과 같다. 먼저 각 진폭을 하나만 사용할 경우의 거리-순서 보존 정도를 구한다. 그런 다음, 거리-순서 보존 정도를 가장 크게 하는 진폭 *f*개를 구하고, 이들 진폭의 인덱스를 구성한다. 이러한 과정을 나타내는 알고리즘이 그림 8(a)의 Algorithm *simple-ddop-index*()이다. 알고리즘을 보면, 먼저 시계열 데이타베이스로부터 임의의 샘플들을 선택한다(라인 (1)). 그런 다음, 각 진폭에 대해서 거리-순서 보존 정도를 계산한다(라인 (2)-(4)). 마지막으로, 거리-순서 보존 정도가 큰 *f*개 진폭의 인덱스를 저장한다(라인 (5)-(7)). 이렇게 저장된 *f*개의 진폭은 단순 탐욕 전략의 각 시계열에서 DFT 진폭을 선택하는데 사용된다. 그림 8(b)는 이러한 단순 탐욕 전략 알고리즘을 나타낸다. 그림 8(b)의 알고리즘은 DFT 진폭을 선택하는데 처음 *f*개 대신 그림 8(a)에서 구한 진폭 인덱스에 의해 *f*개를 선택하는 것(라인 (2))을 제외하고는 그림 7의 순차 선택 전략 알고리즘과 동일하다.

다음으로, 심화 탐욕 전략은 단순 탐욕 전략을 발전시킨 형태로서, 앞서 선택한 진폭을 바탕으로 다음에 선택할 진폭을 결정하는 방식을 취한다. 즉, 첫 번째 DFT 진폭을 탐욕적 방법으로 선택한 후, 두 번째 진폭은 앞서 하나의 진폭을 선택했다는 가정 하에 다시 탐욕적 방법으로 선택하고, 이 과정을 *f*개의 진폭을 선택할 때까지

```
Algorithm simple-ddop-index(a time-series database DB)
(1)  Extract samples of four data sequences from DB;
(2)  for i := 0 to (n-1) do
(3)      Compute ddop[i] from the samples using the i-th DFT magnitude;
(4)  end-for
(5)  for j := 0 to f do
(6)      simple-index[j] := the index of the j-th largest ddop[i];
(7)  end-for
```

(a) 각 진폭의 거리-순서 보존 정도 계산 알고리즘

```
Algorithm simple-greedy-selection (a time-series S of length n, the number f of DFT magnitudes)
(1)  Extract n DFT magnitudes from S by DFT;
(2)  Select the f magnitudes from the DFT magnitudes by the order of simple-index[j];
```

(b) 각 진폭의 거리-순서 보존 정도에 따른 진폭 선택 알고리즘

그림 8 단순 탐욕 전략 알고리즘

```
Algorithm advanced-ddop-index(a time-series database DB)
(1)  Extract samples of four data sequences from DB;
(2)  for j := 0 to f do
(3)      for i := 0 to (n-1) do
(4)          if i is already contained in adv-index[1..(j-1)] then ddop[i] := -1;
(5)          else compute ddop[i] from the samples using the adv-index[1..(j-1)]-th and the i-th DFT magnitude;
(6)      end-for
(7)      adv-index[j] := the index of the largest ddop[i];
(8)  end-for
```

(a) 진폭들의 거리-순서 보존 정도 계산 알고리즘

```
Algorithm simple-greedy-selection (a time-series S of length n, the number f of DFT magnitudes)
(1)  Extract n DFT magnitudes from S by DFT;
(2)  Select the f magnitudes from the DFT magnitudes by the order of simple-index[j];
```

(b) 진폭들의 거리-순서 보존 정도에 따른 진폭 선택 알고리즘

그림 9 심화 탐욕 전략 알고리즘.

반복하는 방식이다. 그림 9(a)의 Algorithm *advanced-ddop-index*()는 이러한 심화 탐욕 전략에 의해 진폭 선택에 사용할 인덱스를 구하는 알고리즘을 나타낸다. 알고리즘의 라인 (3)-(7)을 보면, 앞서 (*j*-1)개의 진폭이 선택된 상태에서 이들 (*j*-1)개의 진폭과 함께 사용되었을 때 거리-순서 보존 정도가 가장 커지는 다음 진폭 하나를 선택한다. 그리고, 이러한 과정을 *f*개의 진폭을 모두 선택할 때까지 반복한다(라인 (2)-(8)). 이런 과정을 통해 저장된 진폭에 대한 인덱스는 각 시계열에서 DFT 진폭을 선택하는 데 사용된다. 그림 9(b)는 이러한 심화 탐욕 전략의 알고리즘을 나타내며, 인덱스로서 *simple-index*[ ] 대신 *adv-index*[ ]를 사용하는 것을 제외하고는 그림 8(b)의 단순 탐욕 전략 알고리즘과 동일하다. 이러한 심화 탐욕 전략은 단순 탐욕 전략에 비해서 계산이 다소 복잡하나, 전역적인 탐욕 접근법을 사용하기 때문에 정확도 측면에서는 보다 우수한 전략이라 할 수 있다.

본 절에서 제안한 세 가지 진폭 선택 전략은 클러스터링 정확도를 높이려는 직관이 적용된 것들이다. 그런데, 이들 세 가지 전략은 어떤 진폭을 사용할지를 선택하는 과정에 있어서 제각기 다른 시간 복잡도를 가진다. 먼저, 순차 선택 전략의 경우 처음 *f*개만 선택하는 단순한 전략으로서, 진폭 선택을 위한 별다른 오버헤드가 존재하지 않는다. 즉, 샘플 수를 *n*이라 했을 때, 유클리디안 거리 계산을 단위 연산으로 하는 시간 복잡도는 $O(1)$이 된다. 다음으로, 단순 탐욕 전략은 각 샘플에 대해서 한 번씩 거리-순서 계산(그림 8(a)에서 *ddop*[*i*]를 계산하는 라인 (3))이 필요하므로, 유클리디안 거리 계산을 단위 연산으로 하는 시간 복잡도는 $O(n)$이 된다. 마지막으로, 심화 탐욕 전략은 *f*개의 진폭 각각을 선택하기 위해서 매번 모든 샘플에 대해서 거리-순서 계산(그림 9(a)에서 *ddop*[*i*]를 계산하는 라인 (5))이 필요하므로, 유클리디안 거리 계산을 단위 연산으로 하는 시간 복잡도는 $O(f \cdot n)$이 된다. 이러한 시간 복잡도 차이에 의

표 1 데이타 집합

번호	종류	시계열 길이	시계열 개수	번호	종류	시계열 길이	시계열 개수
1	50Word	270	454	11	Lighting2	637	60
2	Adiac	176	390	12	Lighting7	324	72
3	Beef	470	30	13	OSULeaf	429	241
4	CBF	128	899	14	OliveOil	571	30
5	Coffee	286	28	15	SwedishLeaf	128	624
6	ECG200	97	99	16	Trace	278	99
7	FaceAll	131	1689	17	Two_pattern	128	3999
8	FaceFour	350	87	18	Synthetic_control	61	300
9	FISH	464	175	19	Wafer	152	6163
10	Gun_Point	152	149	20	yoga	426	2999

한 진폭 선택 과정의 실제 오버헤드는 제6장의 실험 결과에서 확인할 수 있다. 한 가지 유의할 점은 일단 어떤 진폭을 선택할지를 결정한 이후의 클러스터링 시간은 어떤 전략을 사용하든 거의 유사하게 된다는 점이다. 즉, 어떤 진폭을 선택할지 결정하는데 있어서 세 전략의 오버헤드는 다르나, 결정된 이후의 클러스터링 시간에는 큰 영향이 없다는 이야기이다. 결국, 세 전략에 있어서 어떤 진폭을 선택할지를 결정하는 단계는 일종의 전처리 단계로 볼 수 있다.

6. 성능평가

6.1 실험 데이타 및 환경

실험 데이타로는 시계열 대상의 여러 연구[24,25]에서 사용된 UCR 시계열 데이타[26]를 사용하였다. 표 1은 UCR 시계열 데이타의 종류, 시계열 길이, 시계열 개수를 나타낸다. 표에서 보듯이, 각 데이타 집합의 시계열 길이는 61~637로 다양하며, 시계열 개수 역시 30~6163으로 다양하다. 실험에서는 이들 데이타 집합 각각에 대해 거리-순서 보존 정도, 실제 클러스터링 결과, 진폭 선택을 위한 수행 시간을 측정하고 분석하였다.

성능 평가에서는 DFT 계수를 사용하는 기존 전략[8]과 DFT 진폭을 사용하는 세 가지 전략 등 모두 네 가지 방법을 실험하였다. 간략한 표기를 위하여, DFT 계수를 사용하는 기존 전략은 *DFTO*(*DFT original*)로, DFT 진폭을 사용하는 세 가지 전략은 각각 *DFTSS*(순차 선택 전략), *DFTSG*(단순 탐욕 전략), *DFTAG*(심화 탐욕 전략)으로 나타내기로 한다. 실험을 수행한 하드웨어 플랫폼은 Intel Pentium 2.80 GHz, 256MB RAM, 80.0GB 하드 디스크를 장착한 PC이며, 소프트웨어 플랫폼은 GNU/Linux Version 2.6.6 운영체제이다.

6.2 거리-순서 보존 비교 평가

본 절에서는 제안한 세 가지 전략과 *DFTO*의 거리-순서 보존 정도를 비교하는 실험을 수행하였다. 이를 위해, 임의의 시계열 네 개로 구성되는 샘플들을 구축하였는데, 데이타 집합 및 계수(진폭) 개수 각각에 대한 실험을 위해 이러한 샘플들 10,000개를 사용하였다. 즉, 데이타 집합과 계수(진폭)의 개수가 고정된 상황에서, 각 전략에 따른 거리-순서 보존 정도를 10,000번씩 실험하여 그 평균을 결과치로 사용하였다.

먼저, 추출하는 계수(진폭)의 수를 여섯 개로 고정하고, 각 데이타 집합에 따른 거리-순서 보존 정도를 비교하였다. 표 2는 DFT 진폭을 사용하는 *DFTSS*, *DFTSG*, *DFTAG*에서 어느 진폭이 선택되었는지를 나타낸다. 표에서 보듯이, 어떤 전략을 사용하느냐에 따라 선택되는 진폭이 다르게 나타남을 확인할 수 있다. 그림 10은 *DFTO*에 대한 세 가지 전략의 상대적 거리-순서 보존 정도를 나타낸 것이다. 그림을 보면, DFT 진폭을 사용하는 세 가지 전략은 대부분의 시계열 집합에서 DFT 계수를 사용하는 *DFTO*보다는 거리-순서 보존 정도가 낮게 나타남을 알 수 있다. 이는 제안한 세 가지 전략의 경우 위상 정보를 사용하기 않기 때문에 나타나는 당연한 결과로 해석할 수 있다. 여기서 주목할 점은 세 가지 전략과 *DFTO*의 거리 순서 보존 정도의 차이는 평균 약 13%로서 그다지 크지 않다는 것이다. 이는 DFT 계수 기법 대신 DFT 진폭 기법을 클러스터링에 사용하더라도 그 정확도가 크게 나빠지지 않음을 의미한다. 또한, 세 가지 전략 중에서 심화 선택 전략인 *DFTAG*가 가장 좋은 결과를 보임을 알 수 있다. 이는 각각의 진폭뿐 아니라 진폭 사이의 연관성까지 고려하는 심화 탐욕 전략이 진폭 선택을 잘 수행함을 의미한다.

다음으로, 길이가 355인 시계열로 구성된 FaceFour 데이타 집합에 대해 계수(진폭)의 수를 달리하면서 *DFTO*와 제안한 전략들을 비교하였다. 그림 11은 이와 같이 계수(진폭)의 수를 달리한 경우의 *DFTO*에 대한 세 가지 전략의 상대적 거리-순서 보존 정도를 나타낸다. 그림을 보면, 계수(진폭)의 수가 증가함에 따라 제안한 전략들의 정확도가 떨어지는 경우도 있음을 알 수 있다. 이는 위상 정보가 없는 상태에서는 추가적으로 사용되

표 2 UCR 시계열 데이타의 각 데이타 집합에 대한 진폭 선택 결과

시계열 집합	선택된 진폭의 인덱스			시계열 집합	선택된 진폭의 인덱스		
	DFTSS	DFTSG	DFTAG		DFTSS	DFTSG	DFTAG
50Words	0, 1, 2, 3, 4, 5	54, 63, 73, 198, 208, 217	0, 54, 61, 81, 125, 104	Lighting2	0, 1, 2, 3, 4, 5	136, 169, 170, 478, 479, 512	0, 101, 114, 169, 170, 478
Adiac	0, 1, 2, 3, 4, 5	0, 24, 47, 52, 130, 153	0, 30, 45, 47, 54, 57	Lighting7	0, 1, 2, 3, 4, 5	1, 67, 152, 172, 252, 323	0, 1, 33, 64, 291, 323
Beef	0, 1, 2, 3, 4, 5	7, 8, 10, 461, 463, 464	7, 8, 10, 461, 463, 464	OliveOil	0, 1, 2, 3, 4, 5	30, 38, 280, 291, 533, 541	0, 30, 38, 280, 291, 541
CBF	0, 1, 2, 3, 4, 5	0, 21, 50, 54, 75, 79	0, 21, 29, 39, 47, 48	OSULeaf	0, 1, 2, 3, 4, 5	64, 76, 78, 351, 353, 365	0, 64, 66, 76, 87, 363
Coffee	0, 1, 2, 3, 4, 5	0, 1, 134, 142, 153, 286	0, 1, 286, 134, 153, 130	SwedishLeaf	0, 1, 2, 3, 4, 5	0, 2, 53, 63, 65, 126	0, 2, 52, 57, 64, 126
ECG200	0, 1, 2, 3, 4, 5	1, 3, 94, 96, 14, 83	0, 1, 2, 3, 24, 96	synthetic-control	0, 1, 2, 3, 4, 5	0, 1, 11, 19, 50, 60	0, 1, 17, 18, 19, 60
FaceAll	0, 1, 2, 3, 4, 5	0, 36, 39, 42, 90, 96	0, 36, 37, 38, 39, 94	Trace	0, 1, 2, 3, 4, 5	0, 2, 4, 34, 244, 276	0, 2, 31, 34, 244, 276
FaceFour	0, 1, 2, 3, 4, 5	3, 7, 8, 347, 348, 352	0, 3, 7, 8, 347, 348	Two patterns	0, 1, 2, 3, 4, 5	0, 23, 28, 37, 91, 100	0, 14, 21, 31, 36, 55
FISH	0, 1, 2, 3, 4, 5	0, 1, 2, 3, 462, 463	0, 1, 2, 81, 436, 462	wafer	0, 1, 2, 3, 4, 5	0, 5, 6, 71, 146, 147	0, 5, 6, 71, 146, 147
Gun-Point	0, 1, 2, 3, 4, 5	1, 2, 4, 48, 150, 151	0, 1, 2, 4, 150, 151	yoga	0, 1, 2, 3, 4, 5	1, 174, 200, 226, 252, 425	0, 1, 2, 164, 201, 426

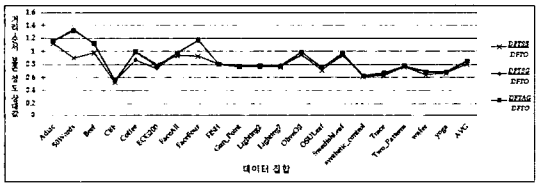


그림 10 UCR 시계열 데이타의 각 데이타 집합에 따른 상대적 거리-순서 보존 정도

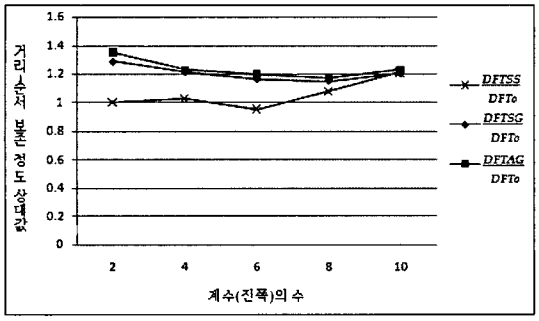


그림 11 계수(진폭)의 수에 따른 상대적 거리-순서 보존 정도의 변화

는 진폭들이 오히려 잡음(noise)으로 작용할 수도 있기 때문이다. 결국, 그림 11의 결과에 따르면 진폭을 많이 사용한다고 정확도가 증가하는 것이 아니므로 각 데이타 집합에 대한 실험을 통해 진폭 개수를 결정하는 것이 바람직하다고 할 수 있다. 그림 11의 결과에서 한 가

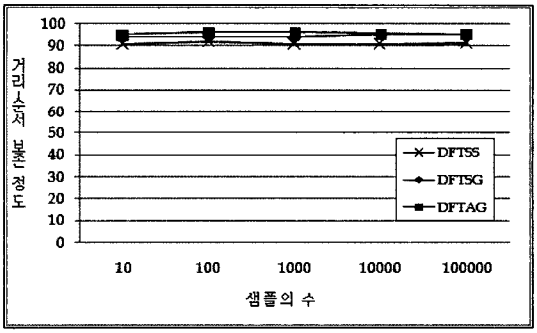


그림 12 샘플의 수에 따른 거리-순서 보존 정도

지 주목할 점은 DFT 진폭의 계수에 관계없이 *DFTAG*가 *DFTSS*나 *DFTSG*보다 우수한 결과를 보이는 것이다. 이는 앞서 그림 10에서도 확인한 바와 같이 탐욕적 전략 중에서도 전역적 접근법이 거리-순서를 가장 잘 보존하는 우수한 전략임을 의미한다.

마지막으로, 샘플 수에 따른 거리-순서 보존 정도를 분석하기 위해 길이가 176인 시계열로 구성된 Adiac 데이타 집합을 대상으로 실험을 수행하였다. 이를 위해, 임의의 시계열 네 개로 구성되는 샘플 집합을 각각 10개, 100개, 1,000개, 10,000개, 100,000개를 구성하였다. 즉, 진폭의 수를 여섯 개로 고정한 후, 각 샘플 집합에 대해 거리-순서 보존 정도와 수행 시간을 측정하였다. 그림 12는 각 샘플 집합에 대한 거리-순서 보존 정도에 대한 실험 결과이다. 그림을 보면, 샘플의 수를 달리하

더라도 거리-순서 보존 정도에는 큰 차이가 없음을 알 수 있다. 이는 작은 수의 샘플을 사용해서도 비교적 정확한 계수 선택이 가능함을 의미한다. 다음으로, 거리-순서 보존 정도는 앞선 실험의 결과와 마찬가지로 심화 탐욕 전략, 단순 탐욕 전략, 순차 선택 전략 순으로 나타났다.

6.3 클러스터링 결과 비교

본 절에서는 제안한 세 가지 전략과 *DFTO*의 실제 클러스터링 결과를 비교하는 실험을 수행하였다. 클러스터링 결과를 비교 평가하기 위한 척도로는 F-measure를 사용한다. F-measure는 정보 검색 시스템의 성능을 평가할 때 사용되는 척도[8,27]로서, 본 연구에서는 이를 클러스터링 결과를 비교 평가하기 위해 사용한다. F-measure는 정확률(precision)과 재현률(recall)에 의해 얻을 수 있는데, 여기서 정확률과 재현률은 각 클러스터마다 계산된다. 정확률은 프라이버시 보호된 데이타를 사용하여 얻은 클러스터 내 객체 중 몇 개의 객체가 원본 데이타를 사용하여 얻은 클러스터 내의 객체와 같은가에 대한 것이고, 재현률은 원본 데이타를 사용하여 얻은 클러스터 내 객체 중 몇 개의 객체가 프라이버시 보호된 데이타를 사용하여 얻은 클러스터 내 객체와 같은가에 대한 것이다. 이러한 정확률과 재현률을 사용하여 식 (6)과 같이 클러스터 C에 대한 F-measure를 계산할 수 있다. 식 (6)에서 *pre*는 정확률로 $\frac{|C \cap C_p|}{C_p}$이고, *rec*는 재현률로써 $\frac{C \cap C_p}{C}$이다. 이때, C는 원본 데이타를 사용하여 얻은 클러스터이고, C_p는 프라이버시 보호된 데이타를 사용하여 얻은 클러스터이다. 전체적인 F-measure는 이렇게 각 클러스터 별로 얻어진 F값을 모두 더하여 계산할 수 있으며, F-measure가 높을수록 정확한 클러스터링을 수행한다고 할 수 있다[8].

$$F = 2 \cdot \frac{pre \cdot rec}{pre + rec} \quad (6)$$

실험에서는 진폭(계수)의 수를 여섯 개로 고정한 다음 *DFTO*와 세 전략들에 대하여 클러스터링을 수행한 후, 각 전략들의 F-measure를 비교하였다. 클러스터링에는 k-means 알고리즘을 사용하였다. 그림 13은 *DFTO*에 대한 세 가지 전략의 상대적인 F-measure를 나타낸다. 그림을 보면, 대부분의 데이타 집합에 대해서 클러스터링 정확도가 제6.2절의 거리-순서 보존 정도(그림 10)와 매우 유사한 결과를 나타냈다. 이는 제안한 거리-순서 보존 정도가 클러스터링 정확도를 잘 반영하고 있음을 의미한다. 또한, 거리-순서 보존 정도의 실험에서와 같이, F-measure를 통한 클러스터링 정확도 역시 심화 탐욕 전략, 단순 탐욕 전략, 순차 선택 전략 순으로 높

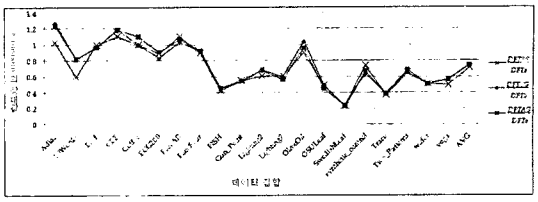


그림 13 UCR 시계열 데이타의 각 데이타 집합에 따른 상대적 F-measure값

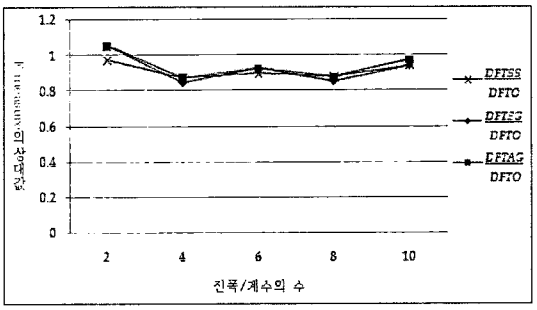


그림 14 계수/진폭의 수에 따른 상대적 F-measure 값의 변화

게 나타났다. 특히, F-measure 측면에서 제안한 심화 탐욕 전략(*DFTAG*)과 DFT 계수 기법(*DFTO*)의 정확도 차이는 평균 20%로 그다지 크지 않은 것으로 나타났다.

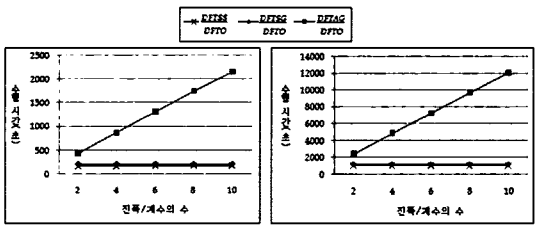
다음으로, FaceFour 데이타 집합에 대해 진폭(계수)의 수를 달리하면서 실험을 수행하였다. 그림 14는 *DFTO*와 세 가지 전략의 상대적인 F-measure값을 나타낸다. 그림 14의 결과는 그림 11의 거리-순서 보존 정도와 매우 유사함을 알 수 있다. 또한, 상대적 F-measure도 그림 11의 거리-순서 보존 정도와 마찬가지로 심화 탐욕 전략, 단순 탐욕 전략, 순차 선택 전략 순으로 높게 나타났다. 이 같은 결과를 종합하면, 제안한 거리-순서 보존 정도가 클러스터링 정확도를 대신할 수 있는 우수한 척도이며, 거리-순서 보존 정도를 사용한 탐욕적 진폭 선택 전략들은 클러스터링 정확도를 높이기 위한 합리적인 기법이라 할 수 있다.

6.4 수행 시간 분석

제5장에서 설명한 바와 같이, 진폭을 선택하기 위한 수행시간 측면에서는 순차 선택 전략과 단순 탐욕 전략이 심화 탐욕 전략에 비해 우수한 결과를 보이게 된다. 본 절에서는 이러한 진폭 선택을 위한 수행 시간을 실험을 통해 확인한다. 그림 15는 두 가지 데이타 집합에 대한 각 전략들의 수행 시간을 나타낸다. 그림 15(a)는 길이가 128인 시계열로 구성된 SwedishLeaf에 대한 실험 결과이고, 그림 15(b)는 길이가 355인 시계열로 구성

된 FaceFour에 대한 실험 결과이다. 그림을 보면, 두 데이타 집합 모두 진폭의 수가 증가할수록 *DFTAG*의 수행 시간이 다른 전략들에 비해 큰 폭으로 증가하는 것을 볼 수 있다. 이는 당연한 결과로서, *DFTAG*의 경우 진폭을 선택할 때마다 모든 샘플에 대해 거리-순서 계산을 수행하기 때문이다. 반면에, *DFTSS*와 *DFTSG*의 경우는 진폭의 수와는 관계 없이 거리-순서 계산을 한번만 수행하므로 수행 시간은 진폭의 수와 관계없이 일정함을 알 수 있다. 특히, 그림 15(b)의 수행시간이 그림 15(a)보다 오래걸리는 이유는 FaceFour의 시계열의 길이가 SwedishLeaf에 비해 약 세 배까지 길기 때문으로 해석할 수 있다.

다음으로, 샘플의 수에 따른 수행 시간을 실험을 통해 확인한다. 이를 위해, 제6.2절에서 설명한바와 같이, Adica 데이타 집합에 대해 진폭의 수를 여섯 개로 고정하고 각각의 샘플 집합에 대해 수행시간을 측정하였다. 그림 16은 샘플 수의 변화에 따른 수행 시간에 대한 실험 결과이다. 실험 결과를 보면, 샘플의 수가 증가 할수록 수행 시간이 증가함을 확인 할 수 있다. 특히, 그림 14의 실험 결과와 마찬가지로 심화 탐욕 전략의 수행 시간이 다른 전략 들에 비해 큰 폭으로 증가함을 알 수 있다. 결국, 심화 탐욕 전략의 경우, 다른 전략들에 비해 클러스터링 정확도는 높지만 진폭 선택 시간의 오버헤드가 큰 것을 확인할 수 있다.



(a) SwedishLeaf의 수행 시간 (b) FaceFour의 수행 시간

그림 15 계수/진폭의 수에 따른 수행 시간의 변화

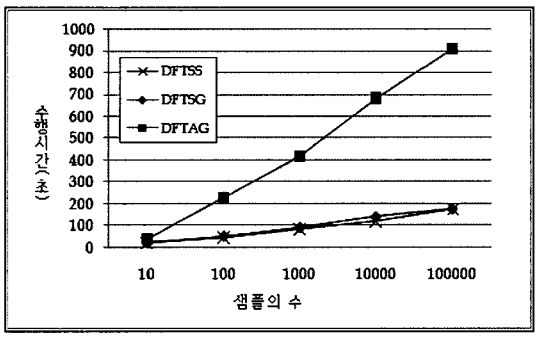


그림 16 샘플의 수에 따른 수행 시간의 변화

7. 결 론

본 논문에서는 시계열 데이타를 대상으로 하는 프라이버시 보호 클러스터링 기법으로 DFT 진폭 기법을 제안하였다. 기존의 DFT 계수 기법[8]은 마이닝 결과의 정확도는 높지만, 프라이버시 보호된 데이타로부터 원본 데이타를 유추(복원)할 수 있어 프라이버시 보호 측면에서 심각한 문제점을 가지고 있다. 이에 따라, 본 논문에서는 프라이버시 보호된 데이타로부터 원본 데이타의 복원이 어려우면서도 마이닝 결과 또한 비교적 정확한 프라이버시 보호 기법을 제안하였다.

본 논문의 공헌은 다음과 같이 요약할 수 있다. 첫째, 원본 데이타의 복원 관점에서 프라이버시 보호 정도를 정의하고, 클러스터링 정확도 측면에서 거리-순서 보존 개념을 정형적으로 제시하였다. 프라이버시 보호 함수가 적용된 데이타에서 원본 데이타의 복원이 어려우면 해당 함수는 프라이버시의 보호 정도가 높은 함수라 설명하였다. 그리고 원본 데이타 내 객체들의 상대적인 순서가 프라이버시 보호 함수가 적용된 후에도 유지되는 정도를 거리-순서 보존 정도라 정의하고, 이러한 거리-순서 보존 정도와 클러스터링 정확도의 관계를 설명하였다. 이에 따라, 본 논문에서는 원본 데이타의 복원이 어려우면서 거리-순서 보존 정도가 높다면 그 함수는 우수한 프라이버시 함수임을 설명하였다. 둘째, DFT 변환 결과에서 위상을 제외한 DFT 진폭만을 사용하는 프라이버시 보호 기법인 DFT 진폭 기법을 제안하였다. 그리고, DFT 진폭 기법이 복원이 어려운 함수임을 체계적으로 주장하였다. 즉, 본 연구에서 제안한 DFT 진폭 기법이 프라이버시 보호 측면에서 우수한 기법임을 설명하였다. 셋째, 클러스터링 정확도를 높이기 위하여 DFT 진폭을 선택하는 순차 선택 전략, 단순 탐욕 전략, 심화 탐욕 전략을 각각 제안하였다. 순차 선택 전략을 제외한 단순 탐욕 전략과 심화 탐욕 전략에서는 거리-순서 보존 정도를 극대화하는 방향으로 진폭을 선택한다. 탐욕적 전략은 진폭 선택을 위한 추가의 오버헤드는 있으나, 추후 클러스터링 정확도를 높이는데 큰 효과가 있는 것으로 확인되었다. 넷째, 제안한 DFT 진폭 기법의 실용성을 다양한 시계열 데이타 집합을 대상으로 하는 실험을 통해 확인하였다. 실험 결과, 우수한 프라이버시 보호 특성을 보이는 DFT 진폭 기법은 거리-순서 보존 정도과 클러스터링 정확도 측면에서도 기존의 DFT 계수 기법에 비교할 만한 결과를 나타내었다. 이 같은 결과를 볼 때, 제안한 DFT 진폭 기법은 기존의 DFT 기반 기법의 프라이버시 보호 문제점을 해결하면서도 비교적 높은 마이닝 정확도를 보이는 우수한 프라이버시 보호 기법이라 사료된다. 또한, 본 연구에서 제안한 탐욕적

계수 선택 기법들을 DFT 계수 기법 등 다른 프라이버시 보호 기법에 적용시키는 연구를 진행할 계획이다.

참 고 문 헌

[1] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," In *Proc. of the Int'l Conf. on Management of Data*, ACM SIGMOD, Dallas, Texas, pp. 439-450, May 2000.

[2] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," *Advances in Cryptology*, Vol. 1807, pp. 35-53, Dec. 2000.

[3] S. Rizvi and J. R. Haritsa, "Maintaining Data Privacy in Association Rule Mining," In *Proc. of the 28th Int'l Conf. on Very Large Data Bases*, Hong Kong, China, pp. 682-693, Sept. 2002.

[4] A. V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," In *Proc. of the 8th Int'l Conf. on Knowledge Discovery and Data Mining*, ACM SIGKDD, Edmonton, Canada, pp. 217-228, July 2002.

[5] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," In *Proc. of the 8th Int'l Conf. on Knowledge Discovery and Data Mining*, ACM SIGKDD, Edmonton, Canada, pp. 639-644, July 2002.

[6] J. Vaidya and C. Clifton, "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data," In *Proc. of the 9th Int'l Conf. on Knowledge Discovery and Data Mining*, ACM SIGKDD, Washington D.C., pp. 24-27, Aug. 2003.

[7] S. R. M. Oliveira and O. R. Zaiane, "Privacy-Preserving Clustering by Object Similarity-Based Representation and Dimensionality Reduction Transformation," In *Workshop on Privacy and Security Aspects of Data Mining*, Houston, Texas, pp. 21-30, Nov. 2004.

[8] S. Mukherjee and Z. Chen, "A Privacy-Preserving Technique for Euclidean Distance-based Mining Algorithms Using Fourier-Related Transforms," *The VLDB Journal*, Vol. 15, No. 4, pp. 293-315, Nov. 2006.

[9] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu, "Time Series Compressibility and Privacy," In *Proc. of the 33th Int'l Conf. on Very Large Data Bases*, Vienna, Austria, pp. 459-470, Sept. 2007.

[10] M. Vlachos, Z. Vagena, P. S. Yu, and V. Athitsos, "Rotation Invariant Indexing of Shapes and Line Drawings," In *Proc. of the Int'l Conf. on Information and Knowledge Management*, Bremen, Germany, pp. 131-138, Oct. 2005.

[11] J. Han and M. Kamber, Data Mining, 2nd Ed., Morgan Kaufmann Publishers, 2006.

[12] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," In *Proc. of the 5th Berkeley Symp. on Math. Stat. Prob.*, California, pp. 281-297, Mar. 1967.

[13] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley-Interscience, 1990.

[14] R. Ng and J. Han, "Efficient and Effective Clustering Method for Spatial Data Mining," In *Proc. of the 20th Int'l Conf. on Very Large Data Bases*, Santiago, Chile, pp. 144-155, Sept. 1994.

[15] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," In *Proc. of the Int'l Conf. on Management of Data*, ACM SIGMOD, Montreal, Canada, pp. 103-114, June 1996.

[16] S. Guha, R. Rastogi, and K. Shim, "A Efficient Clustering Algorithm for Large Databases," In *Proc. of the Int'l Conf. on Management of Data*, ACM SIGMOD, Seattle, Washington, pp. 73-84, June 1998.

[17] E. Keogh, "A Decade of Progress in Indexing and Mining Large Time Series Databases," In *Proc. of the 32th Int'l Conf. on Very Large Data Bases*, A Tutorial, Seoul, Korea, Sept. 2006.

[18] R. Agrawal, C. Faloutsos, and A. N. Swami, "Efficient Similarity Search in Sequence Databases," In *Proc. of the 4th Int'l Conf. on Foundations of Data Organization and Algorithms*, Chicago, Illinois, pp. 69-84, Oct. 1993.

[19] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time-Series Databases." In *Proc. of the Int'l Conf. on Management of Data*, ACM SIGMOD, Minneapolis, Minnesota, pp. 419-429, May 1994.

[20] Y.-S. Moon, K.-Y. Whang, and W.-S. Han, "General Match: A Subsequence Matching Method in Time-Series Databases Based on Generalized Windows," In *Proc. of the Int'l Conf. on Management of Data*, ACM SIGMOD, Madison, Wisconsin, pp. 382-393, June 2002.

[21] E. Keogh, L. Wei, X. Xi, S.-H. Lee, and M. Vlachos, "LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures," In *Proc. of the 32th Int'l Conf. on Very Large Data Bases*, Seoul, Korea, pp. 882-893, Sept. 2006.

[22] Y.-S. Moon, K.-Y. Whang, and W.-K. Loh, "Duality-Based Subsequence Matching in Time-Series Databases," In *Proc. of the 17th Int'l Conf. on Data Engineering*, Heidelberg, Germany, pp. 263-272, Apr. 2001.

[23] G. Bebis, "Image Processing and Interpretation," Lecture Notes. (http://www.cse.unr.edu/~bebis/MathMethods/FT/lecture.pdf)

[24] T. Rath and R. Manmatha, "Word Image Matching Using Dynamic Time Warping," In *Proc. of Computer Vision and Pattern Recognition*, Madison, Wisconsin, pp. 521-527, June 2003.

[25] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, "Fast Time Series Classification Using Numerosity Reduction," In *Proc. of the Int'l Conf. on Machine Learning*, Pittsburgh, Pennsylvania, pp. 1033-1040, June 2006.
[26] E. Keogh, X. Xi, L. Wei, and C. A. Ratanamahatana, The UCR Time Series for Classification/Clustering (http://www.cs.ucr.edu/~eamonn/time_series_data.
[27] F. Crestani, M. Lalmas, C. J. V. Rijsbergen, Information retrieval, Butterworths, 1979.

김 혜 숙
2003년 2월 강원대학교 컴퓨터정보통신공학부 학사. 2006년 2월 강원대학교 교육대학원 컴퓨터교육전공 석사. 2006년 3월~현재 강원대학교 컴퓨터과학전공 박사과정. 2005년 3월~2008년 2월 한국과학기술원 첨단정보기술연구센터 연구원
관심분야는 Data Mining & Knowledge Discovery, Database Security & Privacy, Computer Education

문 양 세
정보과학회논문지 : 데이타베이스
제 35 권 제 2 호 참조