# 시계열 데이터의 프라이버시 보호 클러스터링에서 노이즈 평준화 효과

# (Noise Averaging Effect on Privacy-Preserving Clustering of Time-Series Data)

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

**요 약** 최근, 개인 데이터의 프라이버시 보호에 대한 문제가 대두됨에 따라 대용량 데이터를 대상으로 하는 데이터 마이닝 분야에서도 프라이버시 보호 문제에 대한 활발한 연구가 진행되고 있다. 데이터 마이닝에서의 프라이버시 보호 문제는 정보제공자에 의해 제공된 정보 중 민감한 개인 정보의 노출이 없이도 가능한 정확한 마이닝 결과를 얻는 것이다. 데이터 마이닝의 프라이버시 보호 기법에서는 데이터의 보호뿐만 아니라 결과의 정확도 또한 중요한 요인이다. 이에 따라, 본 논문에서는 시계열 데이터 클러스터링을 기반으로 랜덤 데이터 교란 기법에서 결과의 정확도를 높이는 기법으로 노이즈 평준화 개념을 제시한다. 기존의 랜덤 데이터 교란 기법은 데이터의 프라이버시는 잘 보호하지만 시계열간의 거리-순서가 보존되지 않아 결과의 정확도가 크게 떨어지는 문제점을 가진다. 이를 위해, 본 논문에서는 PAA를 기반으로 하는 노이즈 평준화 개념을 제시하고, 구체적인 예를 통해, 제안한 노이즈 평준화 개념이 랜덤 데이터 교란 기법에서 클러스터링 결과의 정확도를 높일 수 있음을 체계적으로 설명한다.

**키워드** : 시계열 데이터, 프라이버시 보호, 클러스터링

***Abstract*** Recently, there have been many research efforts on privacy-preserving data mining. In privacy-preserving data mining, accuracy preservation of mining results is as important as privacy preservation. Random perturbation privacy-preserving data mining technique is known to well preserve privacy. However, it has a problem that it destroys distance orders among time-series. In this paper, we propose a notion of the noise averaging effect of piecewise aggregate approximation (PAA), which can be preserved the clustering accuracy as high as possible in time-series data clustering. Based on the noise averaging effect, we define the PAA distance in computing distance. And, we show that our PAA distance can alleviate the problem of destroying distance orders in random perturbing time series.

**Key words** : Time-series data, privacy preserving, clustering

## 1. 서 론

최근, 컴퓨터가 처리하는 데이터 양이 증가하고 그 종류가 다양해짐에 따라 개인 데이터의 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 데이터를 대상으로 하는 데이터 마이닝의 여러 분야에서도 활발한 연구가 이루어져왔다[1-9]. 데이터 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 데이터 클러스터링을 기반으로 랜덤 데이터 교란 기법을 다룬다.

본 논문에서는 먼저 랜덤 데이터 교란 기법에서 사용되는 백색 잡음(white noise)의 특성을 분석한 후, 노이즈 평준화 개념을 제시한다. 노이즈 평준화는 백색 잡음의 평균이 0이기 때문에 노이즈들의 합이 0에 가까워진다는 특성에 기반을 둔다. 부분 집계 근사법(Piecewise Aggregation Approximation: PAA)은 시계열 분석에서 주로 사용하는 저차원 변환 기법으로 시계열을 몇 개의 구간으로 나눈 후 그 구간에 대한 평균값을 계산하여 고차원의 시계열을 저차원의 시계열로 변환한다[4]. 본 논문에서는 이러한 백색 잡음의 특성을 기반으로 프라이버시 보호된 시계열들의 거리 계산에 PAA를 적용한다. 즉, 클러스터링을 위해 객체들의 거리를 계산할 때 엔트리 각각에 대한 거리 계산 대신 엔트리들의 묶음의 평균에 대한 거리 계산을 수행한다. 이 개념을 본 논문에서는 PAA 거리(PAA distance)로 정의하였다.

다음으로, 본 논문에서는 클러스터링 정확도를 비교하기 위하여 거리-순서(distance-order)의 개념을 제시한

· 본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다.
· 이 논문은 2009 한국컴퓨터종합학술대회에서 '시계열 데이터의 프라이버시 보호 클러스터링에서 노이즈 평준화 효과'의 제목으로 발표된 논문을 확장한 것임

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

논문접수 : 2009년 8월 14일
심사완료 : 2009년 12월 28일

다[5]. 거리-순서 보존 정도는 원본 데이터에서 객체들의 상대적 순서가 프라이버시 보호된 이후에도 유지되는 정도로써, 이를 사용하여 클러스터링의 정확도를 대신할 수 있다. 마지막으로 기존의 랜덤 데이터 교란 기법인 랜덤화 기법과 웨이블릿 기반의 노이즈 생성 기법에 PAA 거리를 적용하여 PAA 거리의 실용성을 확인한다.

## 2. 관련 연구

Agrawal과 Srikant[1], Lindell과 Pinkas[2]에 의해 데이터 마이닝의 프라이버시 보호 문제가 처음으로 제안된 후, 데이터 마이닝 분야에서도 프라이버시 보호 연구가 많이 시도되고 있다. 최근까지 진행된 데이터 마이닝에서의 프라이버시 보호 연구는 크게 두 분야로 구분할 수 있다. 첫 번째는 데이터 교란을 사용한 프라이버시 보호 기법이다[3,7]. 이 기법은 각 객체의 정보를 보호하기 위해 원본 데이터에 임의의 값 추가(randomization)[3], 원본 데이터의 왜곡(distortion)[6] 등의 전처리 과정을 수행한 후 마이닝을 수행하는 기법이다. 이러한 기존의 데이터 교란 기법은 비교적 간단히 사용할 수 있는 장점이 있으나, 원본 데이터가 하나라도 노출되어 회귀분석(regression)이나 웨이블릿 기반(wavelet-based) 필터링의 과정을 거치면 원본과 유사한 데이터가 복원될 수 있다는 단점이 있다[7]. 이러한 단점을 보완하기 위해 웨이블릿 변환을 사용한 기법이 연구되었다. 이 기법은 노이즈 데이터를 생성하는데 웨이블릿 변환을 사용하는 기법으로 필터링과 과정을 거쳐도 원본 데이터로 복원이 되지 않는 강력한 프라이버시 보호 기법이다[7]. 두 번째는 SMC 기법을 사용한 프라이버시 보호 기법이다[2,8,9]. 이 SMC 기법은 마이닝을 수행할 데이터들이 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 데이터 전송 시 악의적인 사용자에 의해 데이터가 노출될 수 있다는 단점이 있다.

## 3. 노이즈 평준화 효과

### 3.1 노이즈 평준화

랜덤 데이터 교란 기법 중 대표적인 랜덤화 기법(randomization)은 균일 분포(uniform distribution)나 가우시안 분포(Gaussian distribution)를 기반으로 생성한 임의의 값을 원본 데이터에 추가하여 마이닝을 수행한다. 이러한 랜덤화 기법은 프라이버시는 잘 보호하나 마이닝의 정확성은 보장하지 못한다는 단점이 있다[9]. 즉, 랜덤화 기법에서 사용하는 백색 잡음은 프라이버시 보호의 척도가 되는 복원 개념에서 볼 때, 복원이 잘 되지 않는다. 그러나, 랜덤화 기법은 각 객체들의 거리-순서가 잘 보존되지 않기 때문에 클러스터링의 정확성을 보장하기 어렵다. 특히, 프라이버시 보호를 위해 원본 시계열에 노이즈를 많이 더할수록 정확도는 더 떨어진다. 본 논문에서는 이 문제를 해결하기 위해 노이즈 평준화 개념을 제안한다. 랜덤화 기법에서 사용하는 백색 잡음의 평균은 0이기 때문에 이들 노이즈들의 합계는 0에 가깝게 된다. 따라서, 노이즈 평준화는 프라이버시 보호 클러스터링을 수행할 때, 엔트리 각각에 대한 거리 계산 대신에 몇 개의 객체의 평균에 대한 거리 계산을 수행하면 노이즈의 값들이 0에 가까운 값으로 평준화 될 것이라는 개념이다. 즉, 본 논문에서 제안하는 노이즈 평준화 개념은 프라이버시 보호 클러스터링을 수행할 때, 시계열 내의 몇 개 객체의 평균에 대해 거리 계산을 수행하면 클러스터링 정확도가 향상될 것이라는 직관에 기반한다.

본 논문에서는 PAA의 노이즈 평준화 효과를 사용한다. PAA는 시계열분석에서 주로 사용하는 저차원변환 기법으로 계산이 간단하고 성능이 우수하다고 알려져 있다. PAA는 길이가 $n$인 시계열 X(={X[1], …, X[n]})를 다음 식 (1)과 같이 변환한다[4].

$$\overline{x}[i]=\frac{f}{n}\sum_{j=\frac{n}{f}(i-1)+1}^{\frac{n}{f}i}[j],$$

where $f$ is the number of averages. (1)

식 (1)을 사용하여 프라이버시 보호된 시계열 $Xp$(={$X^p$[1], …, $X^p$[$n$]}) 또한 변환할 수 있다. 식 (1)에서와 같이, PAA는 각 구간으로부터 평균값을 계산한다. 이런 PAA를 클러스터링을 위한 거리 계산을 할 때 사용하면 노이즈의 값들이 0으로 평준화되는 노이즈 평준화 효과를 얻을 수 있다. 이를 위해, 본 논문에서는 프라이버시 보호된 두 시계열 사이의 PAA 거리를 다음과 같이 정의한다.

**정의 1.** 프라이버시 보호된 두 시계열 $X^p$와 $Y^p$가 주어졌을 때, 둘 사이의 PAA 거리(*PAA distance*)라고 정의하고, $PD(X^p, Y^p)$는 다음 식을 통해 구해진다.

$$PD(X^p, Y^p)=D(\overline{X^p}, \overline{Y^p})=\sqrt{\sum_{i=1}^{f}(\overline{X^p}[i], \overline{Y^p}[i])^2} \quad (2)$$

원래의 거리 계산 대신 정의 1의 PAA 거리를 사용함으로써, 랜덤 데이터 교란에서의 마이닝 결과의 정확성을 신뢰하지 못하는 문제를 해결할 수 있다.

본 논문에서는 프라이버시 보호 마이닝 기법에서 좀 더 정확한 마이닝 결과를 얻기 위한 방법을 제안하고 있다. 즉, 기존의 프라이버시 보호 마이닝 기법을 적용

한 클러스터링 결과와 본 논문에서 제안한 기법을 적용한 데이터에서 얻은 클러스터링 결과에 대한 비교가 필요하다. 이를 위한 가장 간단한 방법은 원본 및 프라이버시 보호된 데이터에 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 데이터 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 클러스터링 결과의 정확도 대신 사용한다[5].

**정의 2.** 시계열 $O$, $X$, $Y$가 주어졌고, 이들 간 거리를 $D(O,X)$과 $D(O,Y)$라 하자. 또한, 이들 시계열에 프라이버시 보호 함수 $P$를 적용한 결과 시퀀스를 $O^p$, $X^p$, $Y^p$라 하고, 이들 간 거리를 $D(O^p,X^p)$과 $D(O^p,Y^p)$라 하자. 이때, 거리 $D(O,X)$과 $D(O,Y)$의 상대적 순서가 변환 후 거리인 $D(O^p,X^p)$과 $D(O^p,Y^p)$의 상대적 순서와 같다면, $P$는 거리-순서(*distance-order*)를 보존한다고 정의한다. 즉, 다음 식 (3) 또는 (4)가 성립하면, $P$는 거리-순서를 보존한다고 정의한다.

$$D(O,X) \le D(O,Y) \Rightarrow D(O^p,X^p) \le D(O^p,Y^p) \quad (3)$$

$$D(O,X) \ge D(O,Y) \Rightarrow D(O^p,X^p) \ge D(O^p,Y^p) \quad (4)$$

□

다음의 예제 1과 그림 1은 PAA의 노이즈 평준화 효과를 통해 거리-순서가 바뀌는 문제에 대한 예제이다.

**예제 1.** 그림 1에서 보면, 시계열 $O$, $X$, $Y$가 주어졌고, $O$와 $X$의 거리인 $D(O,X)$가 $O$와 $Y$의 거리인 $D(O,Y)$보다 크다고 하자. 각 시계열 마다 20%의 백색 잡음을 더한 후, 거리를 계산했을 때, $D(O^p,X^p)$가 $D(O^p,Y^p)$보다 작아졌다. 즉, 시계열 사이의 거리-순서가 바뀌었다. 이러한 현상은 백색 잡음을 사용하는 랜덤화기법에서 주로 나타나며, 거리 계산을 기반으로 하는 클러스터링 결과의 정확도에 많은 영향을 미친다. 반면, PAA의 노이즈 평준화 효과를 사용하여 PAA 거리 $PD(O^p,X^p)$와 $PD(O^p,Y^p)$를 계산했을 때, $PD(O^p,X^p)$가 $PD(O^p,Y^p)$보다 크다. 이는 PAA 거리가 거리-순서를 잘 보장함을 의미한다.

### 3.2 랜덤 데이터 교란과 PAA 거리

본 장에서는 클러스터링 정확도를 높이기 위하여 PAA 거리를 실제 랜덤 데이터 교란 기법에 적용하는 알고리즘을 제안하고자한다. 즉, 기존의 랜덤화 기법을 그대로 사용하면서 클러스터링을 수행할 때, PAA 거리를 사용하는 것이다. 이러한 과정을 나타내는 알고리즘 randomization_PAA이 그림 2이다. 알고리즘을 보면, 평균이 0이고 표준편차가 σ인 가우시안 분포를 통해 임의의 값을 생성하는 함수 GaussRand(0,σ)를 사용하여 원본 시계열을 왜곡시킨다. 이 결과로 생성된 시계열 $X^p$가 클러스터링에서 사용된다.

그림 3은 그림 2의 알고리즘을 사용한 실제 실험 결과이다. 실험에서는 길이가 2048인 랜덤 워크 데이터를 사용하였다. 각 시계열마다 PAA 특성은 32개를 추출하

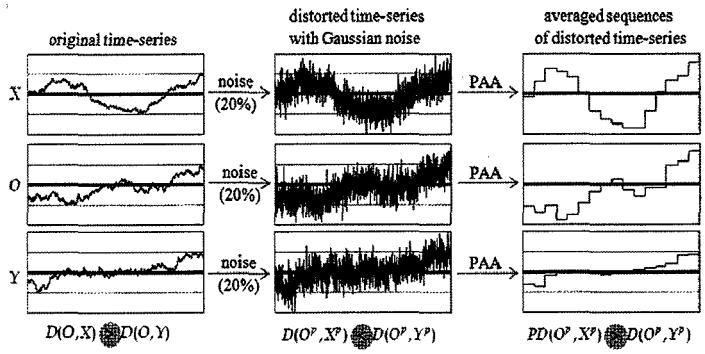


그림 1 PAA 거리와 노이즈 평준화 예제

```
Algorithm randomization_PAA(time-series X of length n, noise σ)
(1) Generate a noise time-seires N(={N[1],...,N[n]}) where N[i]:= Gaussrand(0, σ);
(2) Construct a privacy-preserved time-series Xp from X and N;    // Xp[i] = X[i] + N[i];
(3) Clustering the privacy-preserved time-series Xp;
```

그림 2 Randomization_PAA 알고리즘

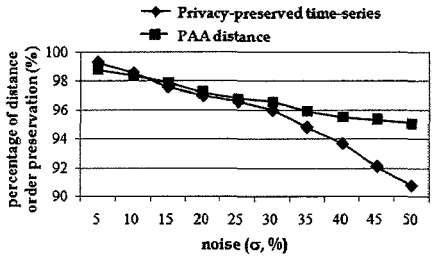


그림 3 Randomization_PAA의 거리-순서 보존 정도

였다. 실험 결과는 PAA 거리의 적용전과 후의 거리-순서 보존 결과를 백분율로 환산하였다. 그림을 보면, PAA 거리를 사용한 경우가 거리-순서를 더 잘 보존하는 것으로 나타났다. 노이즈의 양이 증가할수록 거리-순서 보존정도도 높아지는데, 이는 노이즈의 양이 적을 경우 PAA 거리 자체가 거리-순서에 부정적인 영향을 미친다고 해석할 수 있다. 그러나, 대부분의 경우에 PAA 거리는 노이즈 평준화 효과로 인해 거리-순서 보존에 긍정적인 영향을 미친다.

그러나, 프라이버시 보호 기법으로써 랜덤화 기법은 웨이블릿 필터링을 사용하면 복원이 가능하다는 치명적인 문제점이 있다. 예제 2에서는 이러한 문제점을 나타낸 것이다.

**예제 2.** 그림 4를 보면 원본 시계열에 20%의 노이즈를 더함으로써 프라이버시가 보호된 시계열을 생성했다. 그 결과에 이산 웨이블릿 변환(Discrete wavelet transform: DTW)을 한 후 높은 에너지를 갖는 처음 몇 개의 계수들로 역변환을 하면 원래의 시계열을 복원할 수 있다. 즉, 절대값이 σ미만인 계수들을 제외하고 나머지 계수들을 역변환하는 것이다. 실제 실험결과, 복원된 시계열에는 4.8%의 노이즈만이 남아있었다.

### 3.3 웨이블릿 기반의 노이즈와 PAA 거리

관련연구 [7]에서는 3.2절에서 설명한 바와 같이 랜덤화 기법이 필터링 공격에 약하다는 문제점을 지적하고 새로운 프라이버시 보호 기법으로 웨이블릿 기반 노이즈 생성 기법을 제안하였다. 제안된 웨이블릿 기반 노이즈 생성 기법은 원본 시계열의 웨이블릿 계수를 노이즈 시계열 생성 시에 사용하는 기법이다. 특히, 노이즈 시계열을 생성할 때 높은 에너지를 가지는 웨이블릿 계수만을 사용한다. 그림 5는 웨이블릿 기반 노이즈 생성 기법에서 프라이버시 보호된 시계열을 생성하는 과정이다. 그림을 보면, 먼저 원본 시계열 X를 웨이블릿 변환하여 웨이블릿 계수 $X^w$를 계산한 후, 이를 사용하여 노이즈 시계열을 생성한다. 노이즈 시계열의 생성은 원본의 웨이블릿 계수가 주어진 노이즈의 양인 σ미만일 경우 노이즈 시계열의 계수를 0으로 설정하고, σ이상일 경우에는 노이즈 시계열의 계수를 cσ로 설정한 후, 역변환을 하여 계

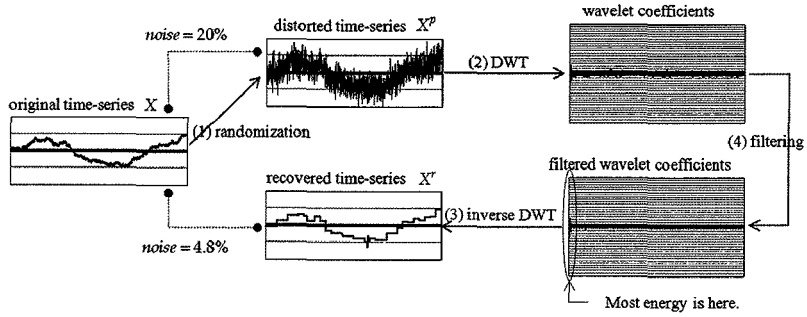


그림 4 랜덤화 기법의 복원 가능 예제

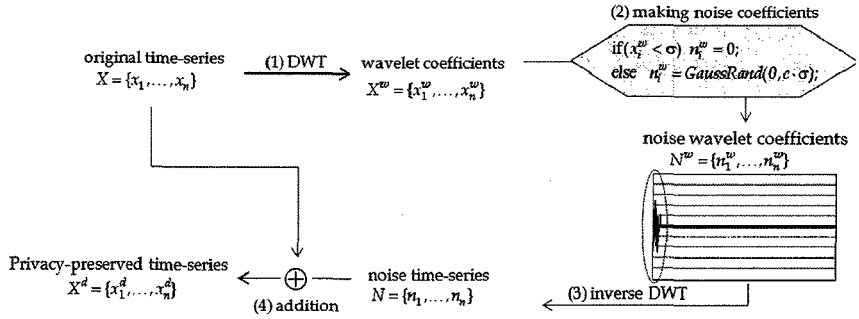


그림 5 웨이블릿 기반의 노이즈 생성 기법

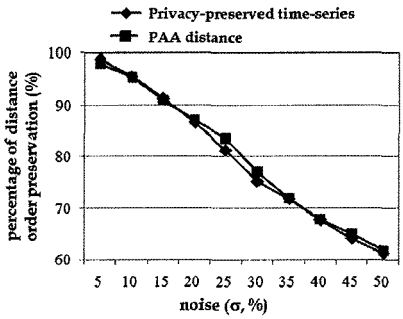


그림 6 Wavelet_PAA의 거리-순서 보존율

산한다. 마지막으로 역변환한 결과인 노이즈 시계열을 원본 시계열에 더해준다. 즉, 원본 시계열의 웨이블릿 계수가 큰 경우에만 노이즈 시계열 생성에 사용한다. 이 기법은 웨이블릿 필터 공격에도 견고하다는 장점이 있다.

본 장에서는 웨이블릿 기반의 노이즈 생성 기법에 PAA 거리를 적용하는 wavelet_PAA 알고리즘을 제안하고자 한다. Wavelet_PAA 알고리즘은 3.2절에서 제안한 randomization_PAA와 같이 기존의 웨이블릿 기반 노이즈 생성 기법에 PAA 거리를 적용한 것이다. 그림 6은 그림 5에서 설명한 절차에 따라 프라이버시 보호된 시계열을 생성한 후, PAA 거리를 적용하여 측정한 wavelet_PAA의 거리-순서 보존 정도에 대한 실험 결과이다. 실험 결과를 보면 PAA 거리를 사용하는 웨이블릿 기반의 노이즈 생성 기법은 PAA 거리를 사용하는 랜덤화 기법에 비해 거리-순서 보존 정도가 급격하게 감소하는 것으로 나타났다. 결과적으로, wavelet_PAA는 높은 에너지를 갖는 몇 개의 계수만 가지고 노이즈를 생성하기 때문에 거리-순서 보존 정도가 떨어지는 것으로 분석된다.

## 4. 결 론

본 논문에서는 시계열 데이터를 대상으로 하는 프라이버시 보호 클러스터링 기법 중 랜덤 데이터 교란 기법의 클러스터링 정확도를 높이는 기법을 제안하였다. 본 논문의 공헌은 다음과 같이 요약할 수 있다. 첫째, 백색 잡음의 특성을 분석하여 노이즈 평준화 개념을 제시하였다. 둘째, 노이즈 평준화 개념을 기반으로 하여 PAA 거리를 제안하였다. PAA 거리는 프라이버시 보호 기법을 적용한 후, 클러스터링을 위한 거리 계산 시에 사용되는 기법으로 기존의 프라이버시 보호 기법을 수정하지 않고 사용할 수 있다는 장점이 있다. 셋째, 클러스터링 정확도 비교를 위해, 프라이버시 보호 기법의 적용 전후 시계열들의 상대적인 유사성을 거리-순서라 정의하였다. 따라서, 본 논문에서는 랜덤 데이터 교란 기법의 거리-순서 보존율을 높이는 기법을 제안하였다. 마지막으로, 본 논문에서 제안한 PAA 거리 기법의 실용성을 실제로 기존의 랜덤 데이터 교란 기법에 적용함으로써 확인하였다. 적용 결과, 랜덤화 기법의 경우, 거리-순서 보존율이 향상되었다. 즉, PAA 거리의 노이즈 평준화 효과를 확인할 수 있었다. PAA 거리를 웨이블릿 기반의 노이즈 생성 기법에 적용한 결과에서는 노이즈의 양과 거리-순서가 상충(trade-off) 관계를 가지는 것으로 분석되었다. 이를 토대로 향후에는 웨이블릿 기반의 노이즈 생성 기법에서 거리-순서의 보존 정도를 높여 클러스터링 결과의 신뢰성을 높이기 위한 연구를 진행하고자 한다.

## 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,* 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] 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*, Edmonton, Canada, pp.217-228, July 2002.

[4] W.-S. Han, J. Lee, Y.-S. Moon, H. Jiang, "Ranked Subsequence Matching in Time-Series Databases," In *Proc. of the 33th Int'l Conf. on Very Large Data Bases,* Vienna, Austria, pp.423-434, Sept. 2007.

[5] H.-S. Kim, Y.-S. Moon, "Privacy-Preserving Clustering on Time-Series Data Using Fourier Magnitudes," *Journal of KIISE: Databases*, vol.35, no.6, pp.481-494, Dec. 2008. (in Korean)

[6] 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.

[7] 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.

[8] 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*, Washington D.C., pp.24-27, Aug. 2003.

[9] 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.