제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

프라이버시를 보존하는 군집화

유현진, 김민호, 라마크리쉬나
광주과학기술원 정보통신공학과
e-mail : hjyoo@gist.ac.kr

Privacy Preserving Clustering

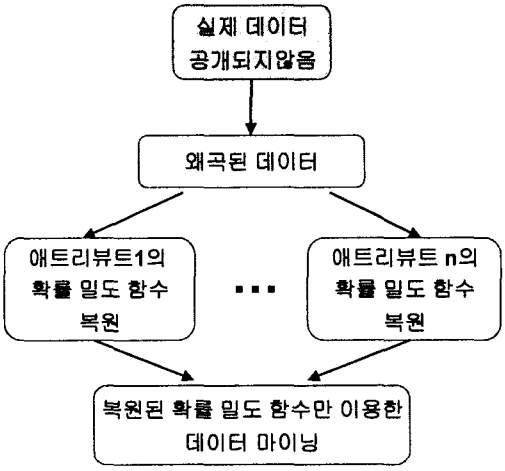
Hyun-Jin Yoo, Min-Ho Kim, R.S. Ramakrishna
Dept. of Information and Communication, Gwangju Institute of Science and Technology

요 약

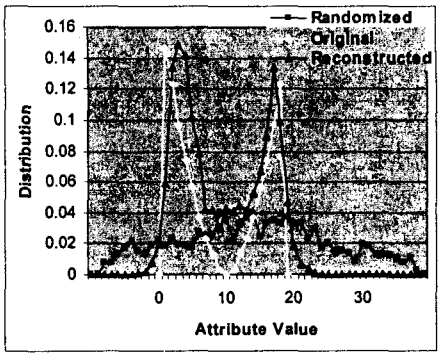
본 논문에서는 프라이버시를 침해 하지 않는 데이터 마이닝에 대해 다룬다. 방대한 데이터에서 유용한 정보를 추출하는 데이터 마이닝분야에서 데이터로부터 프라이버시 보존의 중요성이 부각되고 있다. 그래서 프라이버시의 침해를 막기 위한 방법으로 실제 데이터를 사용하지 않고 잡음이 들어간 데이터를 사용한다. 그리고 프라이버시를 침해하지 않기 위해 잡음이 들어간 데이터로부터 데이터의 확률 밀도 함수(PDF)만을 복원한다. 이렇게 복원된 확률 밀도 함수만을 이용하여 데이터 마이닝기술, 예를 들면 분류화에 곧바로 적용함으로써 프라이버시를 보존하는 것이다. 하지만 분류화에 사용되는 데이터의 1 차원적인 확률 밀도 함수만 가지고는 군집화에 사용하기가 부적절하다. 따라서 본 논문에서는 군집화를 하기 위해 잡음이 들어간 데이터로부터 결합 확률 밀도 함수(Joint PDF)를 복원하고, 복원된 결합 확률 밀도 함수만 가지고 군집화를 할 수 있는 방법을 다룬다.

1. 서론

데이터 마이닝이란, 방대한 데이터에서 묵시적이고 잠재적으로 유용한 정보를 추출하는 작업을 뜻한다. 네트워크의 사용이 지속적으로 증가함에 따라 정보의 디지털화로 인해 산더미 같은 데이터들이 부산물로서 축적되고 있으며, 이들 데이터는 개인에 대한 정보를 담고 있다. 예를 들면, 대형 마트에서는 개인에게 고객카드를 만들게 하여 신상정보와 함께 쇼핑정보를 데이터화 하고, 병원에서는 개인의 병력을 데이터화 하고 있다. 대형 마트에서는 이들 개인의 데이터로 매출의 효과를 높이기 위해 제품의 재배열 등으로 효과를 보고 있으며, 병원에서도 병의 진행상황 파악을 위해 아주 유용하게 쓰이고 있다. 하지만 이러한 데이터들이 공개된다면, 개인의 프라이버시를 침해하게 되는 문제가 발생한다. 일반적으로 프라이버시라 함은 자신에 대한 정보를 남으로부터 보호하는 것을 의미하며, 실제로는 자신의 정보가 오용되지 않는 것이라 할 수 있다. 하지만 문제는 개인의 정보와 관련된 데이터들이 일단 만들어져 공개된다면, 이들 데이터의 오용을 막을 수는 없다는 것이다. 특히, 데이터 마이닝은 이러한 데이터들을 직접 다루어 유용한 정보를 추출하기 때문에, 프라이버시 보존의 중요성이 부각되고 있다. 이러한 프라이버시의 침해를 막기 위한 방법으로 실제 데이터를 사용하지 않고, 잡음이 들어간 변형된 데이터를 사용하는 방법에 관한 연구들이 활발히 진행되고 있으며 프라이버시를 침해하지 않기 위해 잡음이 들어간 데이터로부터 데이터의 확률 밀도 함수(PDF)만을 복원한다[1]. 이렇게 복원된 확률 밀도 함수만을 이용하여 데이터 마이닝 기술, 예를 들면 분류화에 곧바로 적용함으로써 프라이버시를 보존하는 것이다. 하지만 분류화에 사용되는 데이터의 1 차원적인 확률 밀도 함수만 가지고는 동시에 모든 애트리뷰트를 고려하여 결정해야 하는 군집화에 사용하기에는 부적절하다. 왜냐하면, 각 애트리뷰트의 속성을 차례차례 순서대로 고려하는 분류화 기법과는 달리, 군집화에서는 모든 애트리뷰트의 속성을 동시에 고려하여 군집을 결정하기 때문이다. 즉, 1 차원적인 각 애트리뷰트의 확률 밀도 함수들이 아닌, 동시에 모든 애트리뷰트를 고려하는 결합 확률 밀도 함수가 적절하다. 따라서 본 논문에서는 잡음이 들어간 데이터를 사용하



(a) 왜곡된 데이터를 이용한 데이터 마이닝 절차



(b) 데이터의 왜곡과 확률 밀도 복원

그림 1. 복원된 확률 밀도 함수를 이용한 데이터 마이닝

며 이로부터 결합 확률 밀도 함수(Joint PDF)만를 복원하고, 복원된 결합 확률 밀도 함수만 가지고 군집화를 하는 프라이버시를 보존하는 군집화를 다룬다.

본 논문의 2 장에서는 관련연구로서, 관련 연구 및 본 논문의 주요 동기가 되었던 왜곡된 데이터로부터 각 애트리뷰트의 확률 밀도 함수를 복원하여 데이터 마이닝에 적용하는 것을 살펴본다. 3 장과 4 장에서는 각각 프라이버시를 보존하는 군집화를 위해 결합 확률 밀도 함수를 복원해야 하는 필요성과 그 실험 결과를 보이고, 5 장의 결론으로 마무리를 짓는다.

2. 관련연구

프라이버시를 보존하는 데이터 마이닝은 크게 두 가지 방법으로 활발히 연구되고 있다. 하나는 실제 데이터 자체를 왜곡 시키는 것이고[1, 3], 다른 하나는 실제 데이터를 암호화 하는 것이다[2,4,5,6,7].

2.1. SMC (Secure Multiparty Computation)

프라이버시를 보존하는 데이터 마이닝으로 실제 데이터를 사용하는 경우도 있는데, 데이터 마이닝을 할 때 데이터를 암호화하여 다루는 것이다. 이때 데이터를 암호화하여 처리하는 것을 SMC (Secure Multiparty Computation)를 이용한 프라이버시를 보존하는 데이터 마이닝이라 한다. SMC 는 일단 데이터를 암호화한 후, 필요한 계산을 하고, 다시 결과를 복호화하여 반환하기 때문에 이를 이용한 데이터 마이닝은 프라이버시를 보존하게 되는 것이다. 프라이버시를 보존하는 데이터 마이닝에 이용되는 대표적인 SMC 기술로는 Secure Sum, Scalar Product, Secure Set Union, Secure Size of Set Intersection 이 있다.

2.2. 데이터의 왜곡

데이터의 프라이버시를 보존하는 데이터 마이닝을 하기 위해 실제 데이터를 사용하지 않는 방법이 있다. 어떻게 데이터들을 직접 다루어야 하는 데이터 마이닝 기술들을 실제 데이터를 사용하지 않고 가능하게 할 것인가? 그림 1 의 (a)를 보자. 우선, 실제 데이터는 주어지지 않고, 실제 데이터를 왜곡시킨 왜곡된 데이터만이 공개된다. 그래서 이 왜곡된 데이터로부터 각 애트리뷰트의 확률 밀도 함수를 복원하여 데이터 마이닝기술을 적용하는 것이다. 데이터를 왜곡시키기 위해, 실제 데이터에 가우시안 랜덤 분포나 균등 랜덤 분포들을 만들어 더한다. 즉 실제 데이터 값 x_i 대신 x_i+r 을 공개함으로써 데이터의 프라이버시를 보존하는 것이다. 그림 1 의 (b)에서 원본데이터(Original) 와 노이즈가 더해져 왜곡된 데이터(Randomized)를 볼 수 있으며, 왜곡된 데이터의 분포에서 원본데이터의 모습을 전혀 찾아볼 수 없는 점에서 프라이버시가 보존되었음을 확인 할 수 있다.

2.3. 복원된 확률 밀도 함수

왜곡된 데이터와, 데이터를 왜곡시키기 위해 사용된 분포만을 가지고, 각 애트리뷰트의 분포를 복원한다. 다음 수식 1 은 애트리뷰트 X 를 복원하는 공식이며, n 은 애트리뷰트 X 의 값들의 개수이고, f_Y 는 애트리뷰트 X 를 왜곡시키기 위해 더해진 노이즈의 확률 밀도 함수이다. w_i 는 왜곡된 데이터로써, 여기서는 $x_i + y_i$ 이다.

$$f'_X(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X(z) dz}$$

수식 1. 확률 밀도 함수의 복원

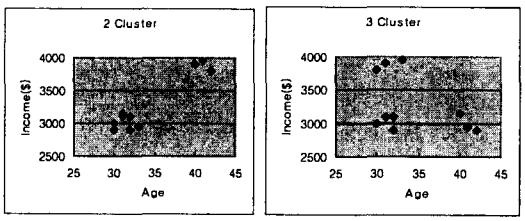
그림 1 의 (b)는 수식 1 을 이용해 데이터의 애트리뷰트 X 를 왜곡시키고, 왜곡된 데이터로부터 복원된 확률 밀도 함수를 보여준다. 본 실험에서는 데이터를 왜곡 시키기 위해 균등 랜덤 분포의 노이즈를 더하였다. 그림에서도 볼 수 있듯이 원본 데이터의 확률 밀도가 거의 유사하게 복원됨을 볼 수 있다.

여기서 주시할 것은 왜곡된 데이터로부터 실제 데이터를 복원하는 것이 아니라, 실제 데이터의 분포만을 복원하기 때문에, 여전히 데이터의 프라이버시가 보존된다는 것이다. 이렇게 복원된 분포를 가지고 데이터 마이닝의 분류화 기법에 적용시킨 것을 논문[1]에서 볼 수 있다.

왜곡된 데이터를 이용하는 데이터 마이닝 기법으로 분류화 기법 외에도 다른 기법들에 대해서도 현재 활발히 연구가 진행되고 있다. 하지만 데이터의 모든 속성을 동시에 고려해야 하는 군집화의 특성으로 인해 프라이버시를 보존하는 군집화에 대해서는 드물게 연구되었다. 본 논문에서는 왜곡된 데이터를 이용하여 프라이버시를 보존하는 군집화 기법에 대해 다룬다.

3. 복원된 결합 확률 밀도 함수를 이용한 군집화

분류화에 사용되는 데이터의 1 차원적인 확률 밀도 함수만 가지고는 군집화에 사용하기가 부적절하다. 그림 2 를 통해 알아보자. 그림 2 의 (a)와 (b)는 각각 군집이 2 개, 3 개인 데이터의 분산형 그래프이다.



(a) (b)

그림 2. 1 차원적인 관점에서
동일한 확률 밀도 함수를 갖지만
서로 다른 군집을 형성하는 데이터

프라이버시를 보존하는 데이터 마이닝에서는 실제 데이터가 아닌 데이터의 확률 밀도 함수만 이용되므로, 데이터 (a)와 (b 각각의 확률 밀도를 살펴보자. 즉, 프라이버시를 보존하는 분류화에서 쓰이는 각 애트리뷰트의 확률 밀도 함수를 보면 (a)와 (b) 모두 다음과 같다.

$$P(25 \le Age < 35) = 0.7$$
$$P(35 \le Age < 45) = 0.3$$
$$P(2500 \le Income < 3500) = 0.7$$
$$P(3500 \le Income < 4500) = 0.3$$

그림 2. (a)와 (b)의 확률 밀도

(a)와 (b)는 클러스터의 개수와 데이터의 분포가 완전히 다름에도 불구하고 각 애트리뷰트의 확률 밀도 함수가 같다. 즉, 각 애트리뷰트의 확률 밀도 함수만 가지고는 (a)와 (b)를 구분 할 수가 없다. 이는 동시에 모든 애트리뷰트를 고려하여 결정해야 하는 군집화의 특성 때문이다. 여기서 (a) 와 (b)의 결합 확률 분포를 보면 다음과 같이 서로 다르다는 것을 알 수 있다.

$$P(25 \le Age < 35 \cap 2500 \le Income < 3500) = 0.7$$
$$P(25 \le Age < 35 \cap 3500 \le Income < 4500) = 0$$
$$P(35 \le Age < 45 \cap 2500 \le Income < 3500) = 0$$
$$P(35 \le Age < 45 \cap 3500 \le Income < 4500) = 0.3$$

그림 2. (a)의 결합 확률 밀도

$$P(25 \le Age < 35 \cap 2500 \le Income < 3500) = 0.4$$
$$P(25 \le Age < 35 \cap 3500 \le Income < 4500) = 0.3$$
$$P(35 \le Age < 45 \cap 2500 \le Income < 3500) = 0$$
$$P(35 \le Age < 45 \cap 3500 \le Income < 4500) = 0.3$$

그림 2. (b)의 결합 확률 밀도

즉, 데이터의 군집화를 위해서는 결합 확률 밀도 함수가 적절하며, 프라이버시를 보존하는 군집화를 하기 위해서 먼저 결합 확률 밀도 함수를 복원해야 한다.

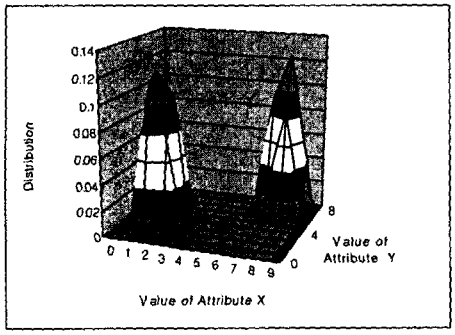
다음공식을 이용해 애트리뷰트가 d 개인 왜곡된 데이터로부터 원본 데이터의 결합 확률 밀도 함수를 복원할 수 있다. $n_1,\cdots,n_d$ 는 각 애트리뷰트의 속성값들의 개수이며, $f_{R_1}(),\cdots,f_{R_d}()$ 은 원본데이터를 왜곡시키기 위해 각 애트리뷰트에 더해진 노이즈들의 확률 밀도 함수들이다. $w_{1i_1},\cdots,w_{di_d}$ 는 왜곡된 데이터로써, $x_{1i_1}+r_{1i_1},\cdots,x_{di_d}+r_{di_d}$ 이다.

$$f_{X_1\cdots X_d}(a_1,\cdots,a_d) = \frac{1}{n_1\cdots n_d}\sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d}\frac{B(i_1,\cdots i_d)}{A(i_1,\cdots i_d)}$$

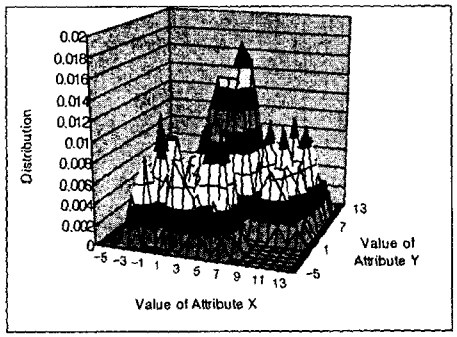
$$A(i_1,\cdots i_d) = f_{R_1}(w_{1i_1}-a_1)\cdots f_{R_d}(w_{di_d}-a_d)f_{X_1\cdots X_d}(a_1,\cdots,a_d)$$

$$B(i_1,\cdots i_d) =$$
$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{R_1}(w_{1i_1}-z_1)\cdots f_{R_d}(w_{di_d}-z_d)f_{X_1\cdots X_d}(z_1,\cdots,z_d)dz_1\cdots dz_d$$

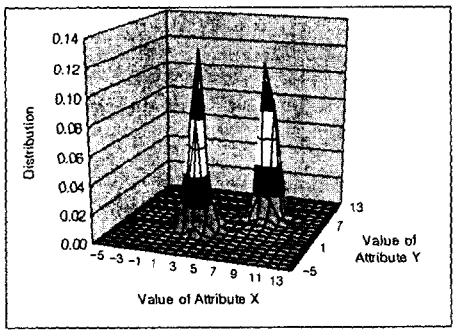
수식 2. 결합 확률 밀도 함수 복원



(a) 원본 데이터



(b) 왜곡된 데이터



(c) 복원된 데이터

그림 3. 결합 확률 밀도

4. 실험결과

결과 확인의 편의를 위해 애트리뷰트가 X, Y 두 개인 데이터를 가지고 실험 하였다. 데이터를 왜곡 시키기 위해 각 애트리뷰트 X, Y에 더해질 노이즈 데이터의 확률 밀도 함수 f_{Rx}, f_{Ry} 는 랜덤 균등 분포를 사용했다. 그림 3 의 (a)는 원본데이터의 결합 확률 밀도 분포이고, (b)는 왜곡된 데이터의 확률 밀도 분포이다. 그림을 통해 알 수 있듯이 그림 3 (b)에서는 원본 데이터의 특징을 전혀 찾아볼 수 없어, 프라이버시가 보존됨을 확인 할 수 있다. 그림 3 (c)는 수식 2 를 이용하여 왜곡된 데이터 (b)로부터 데이터의 결합 확률 밀도 함수를 복원한 것으로써, (a)와 거의 유사하게 복원됐음을 볼 수 있다.

5. 결론 및 향후 과제

본 논문에서는 프라이버시를 보존하기 위해 실제 데이터가 아닌, 왜곡된 데이터로부터 각 애트리뷰트의 확률 밀도 함수를 복원하여 데이터 마이닝에 적용한 방법을 소개하였다. 하지만 각 애트리뷰트의 확률 밀도 함수를 복원하는 것만으로는 군집화에 적용 하기에는 적절하지 않음을 보였다. 즉, 프라이버시를 보존하는 군집화를 하기 위해서는 결합 확률 밀도 함수의 복원이 적절하다는 것을 확인할 수 있었다. 따라서, 실제 데이터가 아닌 왜곡된 데이터로부터 각 애트리뷰트의 확률 밀도 함수가 아닌 결합 확률 밀도 함수를 복원하여 프라이버시를 보존하는 군집화의 가능성을 확인하였다.

복원된 결합 확률 밀도 함수의 실험 결과 그래프에서 확인했듯이, 이미 데이터의 군집들이 결합 확률 밀도 함수에 나타나 있음을 알 수 있다. 결합 확률 밀도 함수를 가지고 군집 정보를 알아내는 것을 향후 과제로 한다.

참고문헌

[1] R. Agrawal and R. Srikant. Privacy - preserving data mining. In Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, pages 439-450, Dallas, TX, May 14-19 2000. ACM.

[2] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. In Proceedings of CRYPTO 2000, LNCS 1880, Springer-Verlag, Santa Barbara, CA, August 2000, pp.36-54.

[3] Stanley Oliveira and Osmar R. Zaiane, Privacy Preserving Clustering By Data Transformation, in Proc. Of the 18th Brazilian Symposium on Databases (SBBD 2003), pp 304-318, Manaus, Brazil 6-8 October 2003.

[4] J. Vaidya and C. Clifton. Privacy Preserving K-Means Clustering over Vertically Partitioned Data. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 2003, pp.206-215.

[5] J. Vaidya and C. Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, July 2002, pp639-644.

[6] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy Preserving Mining of Association Rules. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, July 2002, pp217-228.

[7] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin and M. Y. Zhu. Tools for Privacy Preserving Distributed Data Mining. In SIGKDD Explorations, 4(2): 28-34 Dec. 2002.