다자간 계산과 랜덤화를 복합적으로 사용한 프라이버시 보호 기술에 관한 연구

김종태, 강주성[1)]
국민대학교 자연과학대학수학과
e-mail : jskang@kookmin.ac.kr

A study on the hybrid privacy-preserving techniques by secure multi-party computation and randomization

Jong-Tae Kim, Ju-Sung Kang
Dept. of Mathematics, Kookmin University

요 약

SMC로 불리는 안전한 다자간 계산 프로토콜은 이론적으로 완벽한 프라이버시 보호 기능 및 데이터 정확성을 가지고 있지만 현재의 컴퓨팅 환경에서는 구현이 불가능할 정도로 비효율적이다. 매우 효율적이어서 실용화 되어 있는 랜덤화 기법은 상대적으로 낮은 수준의 프라이버시 보호 기능을 지니고 있다. 최근 SMC와 랜덤화 기법을 적절히 혼합한 형태의 프라이버시 보호 기술이 Teng-Du(2007)에 의해서 제안되었다. 본 논문에서 우리는 Teng-Du의 기법을 면밀히 분석하여 새롭게 구현한 연구 결과를 제시한다. SMC 기술로는 Vaidya-Clifton의 스칼라곱 프로토콜을 채택하고, Agrawal-Jayant-Haritsa가 제안한 랜덤대치 기법을 랜덤화 기술로 선택하여 복합적으로 사용한 프라이버시 보호 기법을 제안한다.

1. 서 론

PPDM(Privacy-Preserving Data Mining)으로 불리는 프라이버시 보존형 데이터 마이닝 기술은 정보 제공자의 중요 정보인 프라이버시를 보호한 상태에서 서버인 마이너가 유용한 정보를 분석해내는 것이다. 프라이버시를 보호하는 기술은 크게 두 가지로 대별된다. 이론적으로 완벽한 안전성과 정확성을 갖는 SMC(Secure Multi-party Computation) 기술은 암호 이론에 기초한 것으로 비트 회로 관점에서 설계가 가능하다. 하지만 대부분의 SMC는 계산 영역이 증가함에 따라서 계산량이 기하급수적으로 팽창하기 때문에 현재의 컴퓨팅 환경에서는 구현이 불가능하다. 한편, 실용적인 프라이버시 보호 기술은 대부분 다양한 랜덤화(randomization) 기법에 의존한다. 하지만 랜덤화를 이용한 프라이버시 보호 기술은 SMC에 비하여 상대적으로 낮은 수준의 프라이버시 보호 기능과 데이터 마이닝 결과에 오차가 발생한다는 약점을 지니고 있다.

이와 같은 SMC 기법과 랜덤화 기법의 특성을 고려하여 최근 이 두 가지 기법을 적절히 혼합한 형태의 프라이버시 보호 기술이 Teng-Du[1]에 의해서 제안되었다. 본 논문에서 우리는 Teng-Du의 기법을 면밀히 분석하고 독자적으로 구현한 연구 결과를 제시한다. 원저자인 Teng-Du는 구체적인 SMC 프로토콜은 언급하지 않았으며, 랜덤화 기법으로는 랜덤응답(randomized response)[2] 프로토콜을 채용하였다. 우리는 현재까지 가장 보편적이고 효율적인 기법으로 알려진 VC 프로토콜[3]과 랜덤대치[4] 기법을 사용한다. 즉, SMC 기술로는 Vaidya-Clifton의 스칼라곱 프로토콜을 채택하고, Agrawal- Jayant-Haritsa가 제안한 랜덤대치 기법을 랜덤화 기술로 선택하여 복합적으로 사용한 프라이버시 보호 기법을 제안한다. 데이터 마이닝 응용 기술로는 분류(classification)를 위한 의사결정나무(decesion tree) 분석 알고리즘 중에서 대표적인 ID3 알고리즘을 고려한다.

2. 프라이버시 보존형 협력 계산 모델

프라이버시 보존형 협력 계산 모델의 효율적인 분류를 위하여 본 연구에서는 ID3알고리즘을 이용하였다. ID3알고리즘의 핵심 아이디어는 다음과 같다.

결정 나무의 말단마디가 아닌 각각의 마디는 입력 속성에 대응되며 각각의 연결은 각 속성의 속성 값과 대응된다. 그리고 출력속성의 기대 값은 뿌리 마디로부터 말단 마디까지의 경로로써 표현되어진다. 그리고 좋은 결정나무가 되기 위하여 각각의 말단마디가 아닌 각각의 마디가 가장 많은 정보를 담고 있을 때 최적의 출력값이 된다.

1948년에 shannon에 의하여 소개된 엔트로피는 불확실성을 측정하기 위하여 사용되었다. 정보이론에서의 엔트로피란 메시지 소스에 대한 불확실성의 측도이다. '메시지 소스의 불확실성이 높을수록 어떤 메시지인지 수신자가 알아내기 위하여 많은 정보를 요구하게 되는가?' 라는 의미에서 정보량을 측정하는 도구로서 엔트로피가 사용 된다.

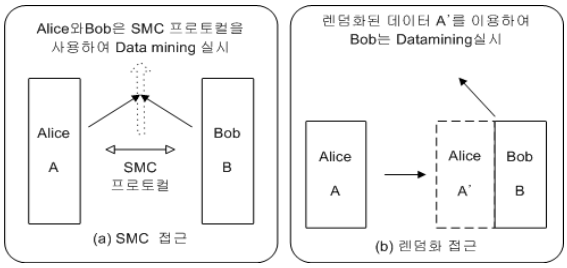
효율적인 분류를 위한 ID3 알고리즘은 주어진 데이터 집합에 대하여 정보이익(information gain)의 값이 높은

1) 교신 저자

속성으로부터 반복적으로 분류를 수행하게 된다. 관심 있는 속성을 나타내는 클래스 속성값이 유일하게 결정되는 노드가 나타나거나, 더 이상 고려해야할 분류 속성이 없을 때까지 ID3 알고리즘은 수행된다.

***ID3*(E, AL) 알고리즘**

1. 마디 V를 생성한다.
2. 만약 마디가 하나의 클래스 속성만을 가지면 말단 노드로 선택한다.
3. 더 이상의 속성 목록 AL이 없으면 가장 많은 클래스 C를 선택하여 말단 마디를 만든다.
4. 가장 큰 정보이익을 갖는 검사 속성 TA를 찾는다.
5. TA를 마디 V라 한다.
6. 각 알려진 TA의 값 a_i 에 대하여
   6.1. $TA=a_i$ 인 마디 V에서 가지를 만든다.
   6.2. $TA=a_i$ 인 데이터 부분 집합 E의 하위 마디 중 하나라도 공집합이면, E의 클래스 값 중 다수인 클래스 C를 선택하여 말단 마디를 만든다.
   6.3. E의 하위 마디 중 클래스가 결정되지 않은 속성값 a_i에 대하여, $ID3(E\wedge(TA=a_i), AL-TA)$를 실행한다.



(그림 1) SMC와 랜덤화에 의한 프라이버시 보호

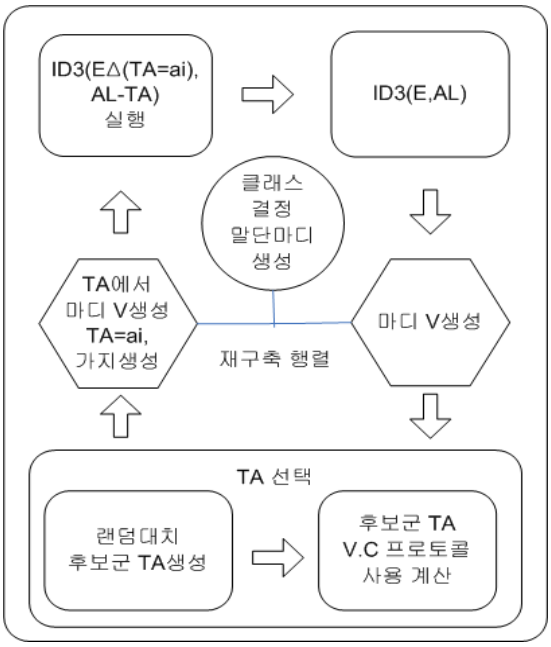
이제 위의 ID3 알고리즘의 프라이버시 보존형 버전을 생각해보자. 문제를 단순화하기 위하여 A,B 두 개체가 존재하고 각 개체의 프라이버시를 보호하는 상태에서 ID3 알고리즘을 통한 데이터 마이닝을 실시한다고 하자. 이 목적을 달성하기 위한 방법으로는 그림 1에 표현되어 있는 SMC만을 이용하는 방법과 랜덤화만을 이용하는 방법, 그리고 그림 2에 나타나 있는 것과 같이 두 가지를 복합적으로 이용하는 방법이 있다. 여기에서 우리는 복합적 기법에 집중한다.

3. 복합-ID3 알고리즘

알고리즘에 참여하는 두 개체는 하나에 대한 어떤 자료를 각각 가지고 서로 상대에게 직접적으로 정보를 노출하지 않으면서 두 개체 모두의 자료로부터 ID3 알고리즘의 결과를 얻고자 한다.

복합-*ID3*(E,AL) 알고리즘

1. 마디 V를 생성한다.
2. 만약 마디가 하나의 클래스 속성만을 가지면 말단 노드로 선택한다.
3. 더 이상의 속성 목록 AL이 없으면 가장 많은 클래스 C를 선택하여 말단 마디를 만든다.
4. 가장 큰 정보이익을 갖는 검사 속성 TA를 찾는다.
   4.1. $D_A\cup\widehat{D_B}=D_1$ 과 $\widehat{D_A}\cup D_B=D_2$ 에서 정보이익 값을 각각 구하고 그 평균값을 구한다.
   4.2. 정보이익 값이 가장 큰 w 개의 검사 속성을 선택한다.
   4.3. 선택한 속성들에 의하여 분할된 TA의 실제의 정보이익 값을 SMC를 사용하여 계산하여 가장 큰 정보이익 값을 가지는 속성을 선택한다.
5. TA를 마디 V라 한다.
6. 각 알려진 TA의 값 a_i 에 대하여
   6.1. 마디 V에서 $TA=a_i$ 인 가지를 만든다.
   6.2. $TA=a_i$ 인 데이터 부분 집합 E의 하위 마디 중 하나라도 공집합이면, E의 클래스 값 중 다수인 클래스 C를 선택하여 말단 마디를 만든다.
   6.3. E의 하위 마디 중 클래스가 결정되지 않은 속성값 a_i에 대하여, $ID3(E\wedge(TA=a_i), AL-TA)$를 실행한다.



(그림 2) 복합-ID3 알고리즘 개념도

복합적 사용기법에서는 랜덤화를 통하여 계산 부하를 줄이면서 SMC를 통해 정확도를 조절한다. 여기서 두 개체는 A의 데이터를 D_A라 하고 B의 데이터를 D_B라 할 때, A는 D_A를 랜덤화한 $\widehat{D_A}$를 B에게 전송 하고 B는 $\widehat{D_A}\|D_B$를 가짐으로써 이 집합에 대한 데이터 마이닝을 실시 할 수 있다. 유사한 방법으로 A도 $D_A\|\widehat{D_B}$를 가지고 데이터 마이닝을 실시할 수 있다.

랜덤대치 기법은 랜덤화 기법 중 구현이 용이하고 응용 가능성이 높다는 장점을 가지고 있으며 프라이버시 손상에 대한 측정이 용이하며, 파라미터에 의하여 프라이버시와 정확도의 취사 선택이 가능하다는 장점을 가지고 있다. 본 실험에서 사용한 랜덤 대치 기법의 기본적인 아이디어는 어떤 확률모델을 사용하여 정의역으로부터 랜덤한 다른 값으로 변화 하는 것이다. 이 확률모델은 각 속성값이 바뀔 확률을 나타내는 전환행렬을 생성하여 정의할 수 있다. 속성의 정의역을 $U=\{u_1,\cdots,u_N\}$라 가정하고 한 데이터의 속성값 u_k가 u_h로 바뀔 확률이라 하면 다음과 같이 정의된다.

$$\Pr[u_k \rightarrow u_h] = m_{h,k} \ .$$

이렇게 정의된 확률값 $m_{h,k}$를 성분으로 하는 $N \times N$ 크기의 행렬을 M이라 하자. Agrawal-Haritsa[5]가 제안한 γ를 사용하여 제시한 최적의 변환행렬은 다음의 형태를 가지는 γ 대각행렬 이다. 변환행렬 $M=xG$는 다음과 같이 정의된다.

$$x=\frac{1}{\gamma+N-1}, \quad G=\begin{bmatrix} \gamma & 1 & \cdots \\ 1 & \gamma & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

본 논문에서는 SMC 프로토콜로 Vaidya와 Clifton이 제안한 양자간 스칼라 곱셈방식[3]을 이용하였다. VC 프로토콜로 명명한 이 프로토콜은 표 1에 나타나 있다.

<표 1> VC 프로토콜

프로토콜 수행과정	
공유 파라미터	▪ 랜덤 $n \times n$행렬 C ▪ 파라미터 $r(<n)$을 A와 B가 공유
step 1 : A	▪ 랜덤 $n-$벡터 $R=(R_1,\cdots,R_n)$생성 ▪ $X'=X+C \cdot R$ 계산
전송 : $A \xrightarrow{X'} B$	
step 2 : B	▪ $S=X' \cdot Y$ 계산 ▪ 랜덤 r벡터 $R'=(R_1,\cdots,R_r)$ 생성 ▪ $Y'=C^T \cdot Y+E(R')$
전송 : $B \xrightarrow{S,\ Y'} A$	
step 3 : A	▪ Temp=$S-R \cdot Y'$ 계산 ▪ R을 이용하여 r벡터 R''생성
전송 : $A \xrightarrow{Temp,R''} B$	
step 4 : B	▪ $X \cdot Y=temp+R'' \cdot R'$ 계산
전송 : $B \xrightarrow{X \cdot Y} A$	

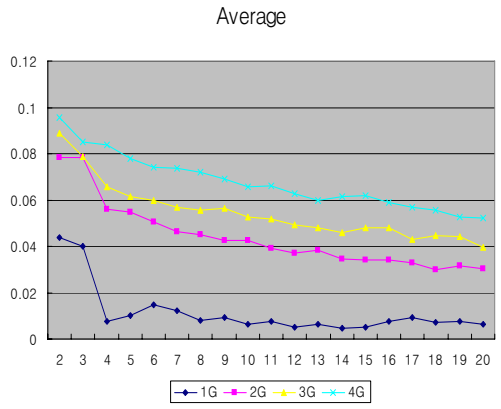
랜덤화 과정에서 감소하는 정확도를 향상시키기 위하여 랜덤화에 사용되는 다중 그룹 구조를 사용한다. 다중 그룹 구조란 속성목록을 $g(1 \leq g \leq t)$개의 그룹으로 나누어 그룹 단위의 랜덤화를 실시하게 된다. 이렇게 하면 그룹 단위로 몇 개의 속성을 하나의 노이즈로 숨기기 때문에 속성의 관계를 좀 더 잘 보전할 수 있게 된다. 일반적으로 랜덤대치 기법에서는 각 속성값을 변환하게 되는데, 이때 다중 그룹 구조에서는 그 변환이 그룹 단위로 이루어지게 된다.

A의 데이터 t개의 속성 목록을 $\{A_1,\cdots,A_t\}$라 하고, 그룹 단위의 속성 목록을 $T_1,\cdots,T_g$라 하자. 그러면 각 그룹의 속성 개수를 $N(T_k)$로 나타낼 때, $t=\sum_{k=1}^{g} N(T_k)$ 가 된다. 랜덤화 작용소는 각각의 그룹 단위로 랜덤화를 적용하게 된다. $E(\cdot)$을 적절한 이진 부호화라고 하면, $E(T_i)$의 값에 대하여 랜덤대치 기법을 적용한다. 즉, 랜덤대치 과정에서 D_A는 각 T_i 그룹 단위로 분할되어 랜덤화가 이루어지는 것이다.

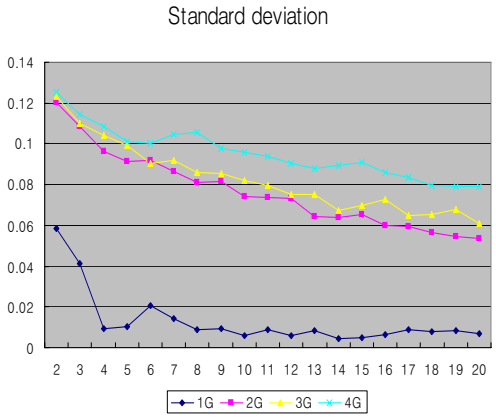
A의 속성 부호화 값을 $E(D_A)=\bigwedge_{j=1}^{g} E(T_j^A)$라 하고, B의 속성 부호화 값을 $E(D_B)=\bigwedge_{j=1}^{g} E(T_j^B)$라고 하면, A와 B의 열벡터를 각각 $X=E(D_A)$, $Y=E(D_B)$로 놓고 VC 프로토콜을 수행한다. 전체적으로는 랜덤대치 기법에 의하여 SMC 계산을 수행할 속성들의 후보를 결정한 후, 그 후보 속성에 대해서만 VC 프로토콜을 수행하기 때문에 SMC만을 적용한 방식에 비하여 계산 효율성이 높아지게 된다.

4. 시뮬레이션 결과

본 실험에서 사용한 데이터는 UCI Machine Learning Repository[7]에 나타나 있는 독버섯 분류 데이터이다. 실험에서 오차는 실제 데이터 정보이익 값과 랜덤화 적용 후의 데이터 정보이익 값의 차이로 측정하였다.



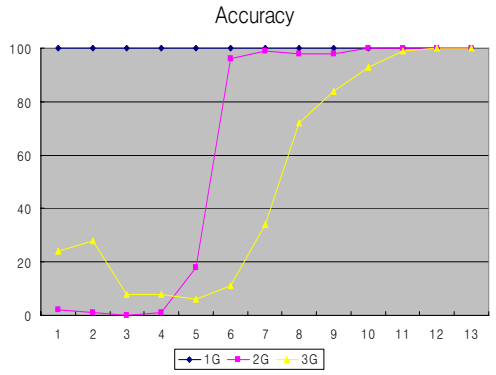
(그림 3) γ 값에 따른 오차 평균



(그림 4) γ 값에 따른 오차의 표준편차.

그림 3은 γ 값의 변화에 따른 정보이익 값 오차의 변화를 보여주고 있다. 랜덤대치 기법 상의 오차는 N이 증가하거나 γ가 1에 가까워질수록 커지게 되는 경향이 나타난다. 우리의 실험 결과에서도 γ 값의 증가에 따라 정보이익 값의 오차 범위가 작아짐을 관찰할 수 있다. 정보이익 값의 범위는 0과 1 사이에서 결정되는데, 오차가 0.1 이하로 관측되므로 어느 정도 만족할만한 결과를 도출 할 수 있으리라 예상된다.

프로토콜의 활용적인 측면에서의 정확도 측정을 위하여 γ에 따른 최대 정보이익 값을 갖는 속성의 적중률을 측정해 보았다. 그림 5에 나타난 결과는 SMC를 수행하는 후보 속성의 개수가 하나인 경우에 실제 정보이익 값이 최대인 속성의 적중률을 측정한 것이다.



(그림 5) γ값에 따른 적중률

후보 속성의 개수를 두 개로 설정 하였을 때, 대부분의 경우 최대의 정보이익 값을 갖는 속성을 찾아내는 것으로 측정되었으며, 후보 속성의 개수를 세 개로 정한 경우에는 100% 최대 속성을 찾아내는 것으로 나타났다. 그러므로 실제 복합적 방법을 적용함에 있어서 후보 속성의 개수를 세 개 이하로 낮춤으로써 원하는 계산 효율성을 얻을 수 있다는 사실을 알 수 있다.

우리는 적절한 γ의 선택으로 랜덤대치 기법을 적용할 경우에 보다 높은 프라이버시 보호 기능을 만족하면서 효율성을 확보할 수 있다. 실험 결과에 의하면 SMC를 적용할 후보 속성의 개수는 세 개 정도면 충분할 것으로 보인다.

5. 결론

본 논문에서는 랜덤화 기법의 사용으로 SMC의 계산 부하를 줄임으로써 효율성을 확보할 수 있는 복합 ID3 프로토콜에 대하여 논하였다. 우리가 다룬 복합적 ID3 알고리즘은 랜덤대치 기법에 의하여 정보획득 값의 계산 대상 속성의 개수를 줄임으로써 효율성을 높인 것이다. 원전에서 사용된 랜덤응답 프로토콜 대신에 랜덤대치 기법을 적용함으로써 정확도가 향상됨을 알 수 있었다.

향후 복합적 기법을 수직 분할 데이터에서 수평 분할 데이터의 경우로 확장하는 메커니즘과 연관규칙 마이닝 등과 같은 다양한 데이터 마이닝 영역으로 일반화 하는 작업은 중요한 연구 과제이다.

참고문헌

[1] Zhouxuan Teng and Wenliang Du, “A Hybrid multi- group privacy-preserving approach for building decision trees”, PAKDD2007, Department of Electrical Engineering and Computer Science Syracuse University, Syracuse, NY 13244, USA.

[2] W. Du and Z. Zhan. "Using randomized response techniques for privacy-preserving data mining". In Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 505-510, Washington, DC, USA, August 24-27 2003.

[3] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in Vertically Partitioned Data", ACM SIGKDD interational conference on knowledge discoveryand data mining, 2002.

[4] S. Agrawal and J. R. Haritsa, "A Framework for high-Accuracy Privacy-preserving Mining", In Proc. of ICDE 2005.

[5] J. Dowd, S. Xu and W. Zhang, "Privacy-preserving Decision Tree Mining Based on Random Substitutions", ETRICS2006, LNCS 3995, Springer-Verlag, pp.145-159, 2006.

[6] A. Evfimievski, J. Gehrke, and R. Srikant, “Limiting Privacy Breaches in Privacy Preserving Data Mining”, *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, 2003.

[7] http://archive.ics.uci.edu/ml/