JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 10, pp. 59-67, October 2020
https://doi.org/10.9708/jksci.2020.25.10.059

Collecting Health Data from Wearable Devices by Leveraging Salient Features in a Privacy-Preserving Manner

Su-Mee Moon*, Jong-Wook Kim*

*Student, Dept. of Computer Science, Sangmyung University, Seoul, Korea
*Professor, Dept. of Computer Science, Sangmyung University, Seoul, Korea

[Abstract]

With the development of wearable devices, individuals' health status can be checked in real time and risks can be predicted. For example, an application has been developed to detect an emergency situation of a patient with heart disease and contact a guardian through analysis of health data such as heart rate and electrocardiogram. However, health data is seriously damaging when it is leaked as it relates to life. Therefore, a method to protect personal information is essential in collecting health data, and this study proposes a method of collecting data while protecting the personal information of the data owner through a LDP(Local Differential Privacy). The previous study introduced a technique of transmitting feature point data rather than all data to a data collector as an algorithm for searching for fixed k feature points. Next, this study will explain how to improve the performance by up to 75% using an algorithm that finds the optimal number of feature points k.

▸ **Key words**: Health Data Collection, Data Privacy, Local Differential Privacy

[요 약]

웨어러블 기기의 발전으로 개인의 건강 상태를 실시간으로 확인하고 위험을 예측할 수 있게 되었다. 예를 들어 심장 질환 환자의 심박수, 심전도가 이상 수치를 보이면 위급 상황을 감지하여 자동으로 보호자에게 연락한다. 이처럼 즉각적인 대처를 가능케 하는 건강 데이터는 생명에 관계되는 만큼 유출되었을 시 심각한 피해를 발생시킨다. 본 연구는 지역 차분 프라이버시 기법을 통해 데이터 소유자의 개인 정보를 보호하면서 데이터를 수집하는 방법을 제안한다. 선행 연구에서는 고정된 k개의 특징 점을 탐색하는 알고리즘으로 전체 데이터가 아닌 특징 점 데이터를 데이터 수집가에게 전송하는 기법을 소개하였다. 이어서 본 연구는 최적의 특징 점 개수 k를 찾는 알고리즘을 이용하여 성능을 최대 75% 향상시키는 방법에 대해 설명할 것이다.

▸ **주제어**: 건강 데이터 수집, 개인정보 보호, 지역 차분 프라이버시

• First Author: Su-Mee Moon, Corresponding Author: Jong-Wook Kim
*Su-Mee Moon (sumeedi@naver.com), Dept. of Computer Science, Sangmyung University
*Jong-Wook Kim (jkim@smu.ac.kr), Dept. of Computer Science, Sangmyung University
• Received: 2020. 08. 11, Revised: 2020. 09. 20, Accepted: 2020. 09. 20.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 IoT 기술과 웨어러블 기기의 발전으로 스마트 워치, 스마트 밴드와 같은 소형기기에 다양한 센서를 탑재할 수 있게 되었다. 예를 들어 스마트 워치에는 고도를 측정할 수 있는 기압고도계, 심장의 전기적 활동 상태를 나타내는 심전도 센서, 움직임을 추적하는 가속도계 등이 내장되어 있다. 웨어러블 기기는 내장된 센서를 통해 착용자에게 실시간으로 외부 환경 정보 및 건강 상태를 제공한다. 또한 해당 정보는 클라우드 또는 기기 내부에 저장되어 착용자의 건강 상태 또는 운동 활동 모니터링에 이용된다. 이러한 정보는 분석 기술을 통해 질병 조기 진단, 위급 상태 알람 등 다양한 방식으로 활용되고 있다. 예를 들어 심전도 센서를 통해 수집된 데이터는 부정맥, 심방 심실의 비대, 폐순환 장애 등의 질병 진단에 사용된다. 이처럼 웨어러블 기기를 이용한 건강 상태 분석 기술은 현대 빠른 인구 증가와 고령 인구 증가에 대한 해결책을 제시해 준다. 예를 들어 주기적인 건강 상태 확인이 필요한 경우, 웨어러블 기기를 통해 이상 신호를 미리 감지하고 최적의 건강 상태를 유지할 수 있다 [1]. 웨어러블 기기를 이용한 헬스 케어 시스템은 비용이 적게 들고 시공간적인 제약이 없기 때문에 활발하게 연구되고 있는 분야이다. 대표적으로 'EpiWatch'는 가속도계와 심박 센서를 통해 착용자의 발작을 예측하고 가족이나 보호자에게 자동으로 연락하는 어플리케이션이다.

웨어러블 기기를 통해 수집한 건강 데이터는 위와 같은 개인적인 목적 외에도 다수로부터 수집하여 질병 일반에 대한 연구 목적으로 사용 된다 [2]. 이를 위해 다수로부터 데이터를 수집할 수 있는 도구와 건강 데이터 분석을 위한 시스템 등이 개발되고 있다. 예를 들어 애플은 2015년 전 세계 건강 데이터를 수집하는 소프트웨어인 리서치킷을 발표하였다. 해당 소프트웨어를 통해 의사, 과학자 및 연구자는 연구를 위한 의료 빅데이터를 수집할 수 있게 되었다. Chan et al. [3]은 리서치킷을 이용하여 전 세계의 천식 환자들로부터 위치 데이터와 공기 오염도, 증상을 확인하기 위한 설문 등에 대한 데이터를 수집하였다. 이렇게 수집된 건강 데이터는 각자의 연구 목적에 알맞은 방식으로 활용된다. 머신 러닝 기법을 통해 질병을 예측하거나 질병 발생 요인 등을 분석할 수 있다. Wang et al. [4]은 환자들의 의료 기록에서 특징을 추출하여 질병을 분류하고 질병 간 상관관계를 파악하는 모델을 구축했다. 이처럼 건강 데이터는 개인의 건강을 위한 모니터링 목적 외에도, 빅데이터로 활발하게 이용되고 있다.

동시에 건강 데이터는 사생활이 침해될 수 있는 민감한 데이터다. 심전도, 혈압 등의 데이터로 건강 상태 또는 환자가 앓고 있는 질병까지 추측할 수 있기 때문이다. 해당 정보가 제3자에 의해 오·남용될 시, 개인을 죽음에까지 이르게 할 수 있다. 예를 들어 환자에게 약물을 투여하는 의료기기의 포도당 모니터 신호가 탈취된다면 큰 사회적 문제로 발전할 수 있다. 이처럼 건강 데이터 유출에 내재된 위험성에 의한 사고를 방지하고, 적절하게 수집 및 활용하기 위해서 건강 데이터를 보호하는 기법이 필요하다. 특히 최신의 웨어러블 기기는 다양한 건강 데이터를 수집하고 있기 때문에 개인 정보 보호 방안이 필수적이다.

데이터를 안전하게 수집하기 위한 방법으로 중 하나로 지역 차분 프라이버시가 존재한다. 지역 차분 프라이버시는 신뢰할 수 있는 데이터 수집자 없이도 데이터를 수집할 수 있으며 구글, 삼성, 애플 등의 글로벌 기업이 채택하여 사용하고 있는 방법이다. 웨어러블 기기를 통해 건강 데이터를 수집할 때 지역 차분 프라이버시를 이용하면 사생활 침해 없는 수집이 가능하다. 웨어러블 기기 착용자, 즉 데이터 소유자는 본인의 건강 데이터에 노이즈를 추가하여 수집가에게 변조된 데이터를 전송한다. 데이터를 전달 받는 수집가 측면에서는 유출에 대한 부담을 완화할 수 있으며 데이터 소유자 측면에서는 사생활 침해에 대한 위험도가 줄어들어 데이터 제공에 대한 심리적 장벽을 낮출 수 있다. 즉 지역 차분 프라이버시를 통해 서비스 제공자는 다수의 데이터를 수집할 수 있게 되어 고품질의 서비스를 제공할 수 있고, 데이터 소유자는 해당 서비스를 통해 편리함을 얻게 된다.

본 연구는 웨어러블 기기에서 사생활 침해 없이 건강 데이터를 수집하는 방법에 대해 제안한다. 지역 차분 프라이버시를 사용하여 데이터 소유자 측면에서 변조한 데이터를 데이터 수집가에게 전송한다. 제안하는 특징 점 추출 알고리즘을 통해 효율적으로 건강 데이터를 수집할 수 있다. 본 연구의 구성은 다음과 같다. 첫 번째 관련 연구에서는 건강 데이터를 활용하는 방법에 대한 선행 연구를 살펴본다. 두 번째 배경 및 문제 정의에서는 지역 차분 프라이버시 기법을 설명한 후 실험에서 사용한 데이터와 기법 등을 설명한다. 세 번째 본문은 제안하는 특징 점 추출 과정의 구조와 알고리즘에 대해 설명한 후, 실제 걸음 수 데이터에 적용한 모습을 보인다. 마지막으로 결론에서는 실험 결과와 시사점에 대해 언급할 것이다.

II. Related Work

웨어러블 기기를 통해 건강 데이터를 실시간으로 측정할 수 있게 되었다. 병원에서 의료 기기로 몸을 측정하고

질병을 진단하던 과거와 달리, 현재는 웨어러블 기기를 이용하여 집에서 간편하게 건강 상태를 확인하고 질병을 예방할 수 있다. 웨어러블 기기와 건강 데이터 분석 기술의 발전은 헬스 케어 분야의 패러다임을 변화시켰다. 예를 들어 Wu et al. [6]은 EHR(Electronic Health Records)과 -Omic 데이터(유전체학, 단백질 유전 정보학 등) 분석을 통해 개인 맞춤형 의료 서비스가 가능함을 주장했으며 Manogaran et al. [7]은 웨어러블 센서를 통해 측정한 혈압, 혈당치, 심박 수 데이터로 환자의 심장 질병 상태를 예측하는 연구에서 81.52%의 정확도를 보였다. 이처럼 빅데이터는 헬스 케어 분야에서 새로운 비즈니스 모델을 구축하였다. 작업 방식의 변화로 인해 생산성이 향상되고 비용이 절감하였으며, 고객 만족도가 증가하였다 [5].

각종 센서가 탑재된 웨어러블 기기의 등장으로 측정할 수 있는 건강 데이터는 다양화되고 있다. 눈물을 통해 포도당을 측정하여 혈당치를 알려주는 스마트 렌즈, 심전도가 비정상적인 모습을 보일 때 알림을 보내는 시스템인 iHeart, 팔에 연결된 장치로 혈압을 측정하는 벨트는 실시간으로 건강 상태를 측정하기 위해 개발된 웨어러블 기기다. 해당 기기를 착용하면, 병원 외의 환경에서도 건강 상태를 모니터링할 수 있다. 이러한 웨어러블 기기들은 일상에서 사용하기에 부피가 크고, 측정할 수 있는 건강 데이터도 제한적이기 때문에 주로 당뇨병, 심장 질환자 등 환자를 위해 사용되었다. 하지만 최근 센서들이 소형화되고, 스마트 워치와 스마트 밴드처럼 간편하게 일상적으로 착용할 수 있는 웨어러블 기기가 등장하면서 적용 대상이 환자에서 일반인으로 확대되었다. 즉 일반인을 대상으로 건강 데이터를 수집하여 발생 가능한 질병을 예측하고 예방할 수 있게 되었다. 최신 스마트 워치는 심박 수 측정 센서, 가속도계, 자이로스코프 센서, 광 센서 등이 내장되어 있다. 센서에서 측정된 정보는 알고리즘을 통해 유의미한 정보로 가공된다. 예를 들어 가속도계를 통해 걸음 수와 소비한 칼로리를, 심박 수를 통해 수면 상태를 짐작해낼 수 있다. 가공된 데이터는 스마트 워치 착용자에게 제공된다. 이처럼 센서 종류 및 개수, 가공 방식에 따라 다양한 건강 데이터를 수집할 수 있다.

스마트 워치로 수집한 건강 데이터는 심장 질환 환자의 갑작스러운 상태 악화에 대비하거나, 착용자의 비정상적인 움직임, 응급 상황을 감지하는 데 활용된다. 예를 들어 Chaudhuri et al. [8]은 가속도계, 자기계, 자이로스코프 센서를 사용하여 노인을 위한 넘어짐 감지 장치를 개발하였다. 만약 넘어짐을 감지하면 알람을 통해 주변 사람에게 도움을 청한다. 해당 장치의 민감도는 94.1%와 94.4% 사이고, 특정성은 92.1%와 94.6%로 높은 정확도를 보였다. 마찬가지로 가속도계와 자이로스코프 센서를 통해 걸음걸이를 진단하는 방법, 공사 현장에서 노동자의 움직임 모니터링을 통해 몸(어깨, 팔꿈치 등)과 연관된 위험성을 측정하는 연구가 진행되었다 [9], [10].

웨어러블 기기를 통해 간편하게 건강 데이터를 수집하는 방법과 동시에 해당 민감 데이터를 보호하는 방법에 대한 연구도 활발히 수행되고 있다. 그중 차분 프라이버시는 의료 데이터를 공격자로부터 보호하는 최신의 방법이다. 예컨대 Beaulieu-Jones et al. [11]는 의료 데이터에 차분 프라이버시를 적용하여 효율적으로 프라이버시 보존 딥 러닝 모델을 학습시켰으며 Guan et al. [12]은 의료 IoT 기기에서 추출한 데이터에 차분 프라이버시와 머신 러닝 기법을 결합하여 K-means 군집화를 수행하였다. Mohammed et al. [13]는 라플라스 매커니즘과 차분 프라이버시로 개인 정보 유출 없이 암 환자를 데이터 마이닝하는 방법을 제안하였다. 이 외에도 차분 프라이버시는 다른 암호화 기법과 함께 사용되기도 한다. Tang et al. [14]은 차분 프라이버시와 BGN(Boneh-Goh-Nissim cryptosystem), SSS(Shamir's Secret Sharing)를 통합하여 데이터 수집 시 개인 정보를 보호하기 위한 방법으로 사용하였다.

III. Background and Problem Statement

1. Local Differential Privacy

차분 프라이버시는 배경 지식을 지닌 공격자가 변조된 통계치를 통해 특정 개인을 추정할 수 없도록 보장해주는 프라이버시 보호 모델이다. 다수의 개인이 정보를 데이터 수집가에게 보내면, 데이터 수집가는 임의화 알고리즘을 통해 변조된 통계치를 배포한다. 이 때 데이터 수집가는 신뢰할 수 있다고 가정되며 다수의 원본 데이터를 지닌다. 반면에 지역 차분 프라이버시는 신뢰할 수 있는 데이터 수집가를 가정하지 않으며, 데이터 수집가는 원본 데이터가 아닌 변조된 데이터를 전송 받게 된다. 지역 차분 프라이버시에서 각각의 데이터 소유자는 지역적으로 차분 프라이버시를 만족하는 임의화 알고리즘 M을 통해 데이터를 변형한 후 데이터 수집가에게 전송한다. 두 개의 이웃한 데이터베이스 D와 D'이 하나의 레코드만 다를 때, 다음과 같은 수식을 만족한다. 즉 공격자는 높은 확률로 D와 D'를 구분할 수 없으며, 프라이버시 보존 정도는 ϵ으로 조절된다 [15, 16].

$$\frac{\Pr[M(D)=O]}{\Pr[M(D')=O]} \leq e^{\varepsilon}$$

지역 차분 프라이버시에서 ϵ은 sequential composability에 의해 n개로 나누어 사용할 수 있으며, 각각의 데이터는 ϵ_i $(0 \le i \le n)$와 임의화 알고리즘을 통해 차분 프라이버시를 만족하며 변조된다 [17, 18]. 프라이버시 강도와 데이터 유용성은 트레이드 오프 관계이므로 ϵ은 프라이버시 수준과 동시에 데이터 활용도를 결정한다. 즉 ϵ이 작을수록 프라이버시 강도가 높아지는 반면 데이터 유용성은 낮아진다. 따라서 데이터 수집가 측면에서 적절한 ϵ을 선택해야 한다.

2. Notation and Problem Statement

본 논문은 웨어러블 기기에서 개인 정보 침해 없이 건강 데이터를 수집하는 방법을 제안한다. 효율적으로 ε을 사용하여 데이터 유용성을 높이기 위해, 소유자는 전체 데이터가 아닌 일부 특징 점 데이터를 수집가에게 보낸다. 최선의 특징 점을 선택하는 방법에 대해 설명하기에 앞서, 본 섹션은 사용한 수식에 대해 명시한다.

실험에서 수집한 총 소유자 수를 w라고 할 때, 소유자 집단을$U=\{u_1,u_2,\cdots,u_w\}$과 같이 나타낸다. 그리고 n시간 측정한 i번째 소유자 u_i의 시계열 건강 데이터를 $u_i=\left(\left(t_1,x_1^i\right),\left(t_2,x_2^i\right),\cdots,\left(t_n,x_n^i\right)\right)$로 표기한다. 이때 u_i는 단조 증가 형태이므로 $x_1^i \le x_2^i \le \cdots \le x_n^i$를 만족한다. 데이터 소유자는 u_i에 지역 차분 프라이버시를 적용하여, $pu_i=\left(\left(t_1,px_1^i\right),\left(t_2,px_2^i\right),\cdots,\left(t_n,px_n^i\right)\right)$를 수집가에게 전송한다. 라플라스 매커니즘을 통해 노이즈를 추가하며 변조된 데이터는 다음과 같다.

$$px_r^i = x_r^i + Lap\left(\frac{\Delta s}{\varepsilon/n}\right)$$

전역 민감도 Δs는 $x_{\max}-x_{\min}$이고 ε은 프라이버시 강도이다. U에게 수집한 변조된 데이터 소유자 집단을 $PL=\{pu_1,pu_2,\cdots,pu_w\}$라고 할 때, 시간대별 평균은 다음과 같다.

$$AVG(px_r^i)=\frac{1}{w}\times\sum_{pu_i\in PL}^{w}px_r^i$$

IV. Proposed Method to Collect Health Data from Smartband Users

제안하는 프라이버시 보존 데이터 수집 방법은 데이터 소유자 측면과 데이터 수집가 측면으로 분류된다. 데이터 소유자 측면은 측정한 건강 데이터의 특징 점의 위치와 개수를 찾은 후 해당 데이터를 데이터 수집가에게 전송하는 과정이다. 만약 측정한 모든 데이터 u_i에 지역 차분 프라이버시를 적용하여 전송한다면 $\frac{\epsilon}{n}$만큼의 노이즈가 추가되는데, 이는 데이터 유용성이 매우 낮아짐을 뜻하므로 전체 데이터가 아닌 일부 데이터를 전송하는 방법이 필요하다. 선행 연구에서는 k를 임의로 설정했을 때, 최적의 특징 점을 찾는 알고리즘에 대해 제안하였다 [19]. 본 논문은 이전 연구의 연장선상에서 최적의 특징 점 개수인 k를 탐색하는 알고리즘을 통해 보다 효율적으로 건강 데이터를 수집하는 방법을 명시할 것이다. 즉 탐색한 특징 점으로 추정한 회귀 모델의 적합도를 R^2로 측정하여 최적의 특징 점 개수 k를 찾아 데이터 수집가에게 전송하는 방식이다. 그리고 데이터 수집가 측면에서는 전송받은 변조된 특징 점을 통해 $AVG(x'_i)$를 계산하는 방법에 대해 설명할 것이다.

1. Data Owner's Device-side Processing

데이터 소유자는 데이터 유용성을 최대화할 수 있는 특징 점 개수와 위치를 탐색하여 데이터 수집가에게 전송하는 것을 목표로 한다. Fig. 1은 전체적인 프로세스를 나타내며 Fig. 2는 최적의 특징 점 개수를 찾는 알고리즘의 의사 코드이다. k의 범위만큼 Fig. 3을 호출하여 특징 점의 위치를 찾아 회귀 모델을 구축하고 적합도를 R^2로 계산한다. 마지막으로 가장 높은 R^2를 보이는 k와 C_{list}를 찾아 반환한다 (lines 4 – 11).

Fig. 3은 Fig. 1에서 호출되는 특징 점 위치 탐색 알고리즘의 의사 코드이다. 완전 탐색 알고리즘으로, 특징 점을 직선으로 이었을 때 제곱 오차를 최소화하는 위치를 반환 한다. Fig. 3에 명시되어 있는 Step 1은 선행 연구 [16]에서 설명하였으며 Step 2는 Step 1에서 반환된 C_{list}로 회귀 모델을 세우고 R^2를 계산하는 과정이다. 우선 탐색한 특징 점 (t_r,x_r^i)에 노이즈를 추가한다. 이때 노이즈는 라플라스 분포에서 추출되는데, 라플라스 분포는 ε을 ε/k로, Δs를 $x_{\max}-x_{\min}$로 설정한 값을 따르므로 k에 따라 서로 다른 분포를 가진다 (line 22). 해당되는 분포에서 무작위로 선택될 노이즈를 추정하기 위해 데이터 개수만큼 노이즈를 무작위 추출한 뒤 더하여 r을 구한다 (line 23). 이때 무작위 추출되는 노이즈는 역누적분포함수를 통해 0.25와 0.75 사이의 확률을 갖는 값을 무작위 추출하여 사용하였다. 예를 들어 데이터의 개수가 290개라면, 라플라스 분포에서 290번 무작위 수를 추출하여 합을 구한다. 그리고 r을 C_{list}의 각 특징 점에 더하여 노이즈가 추가된

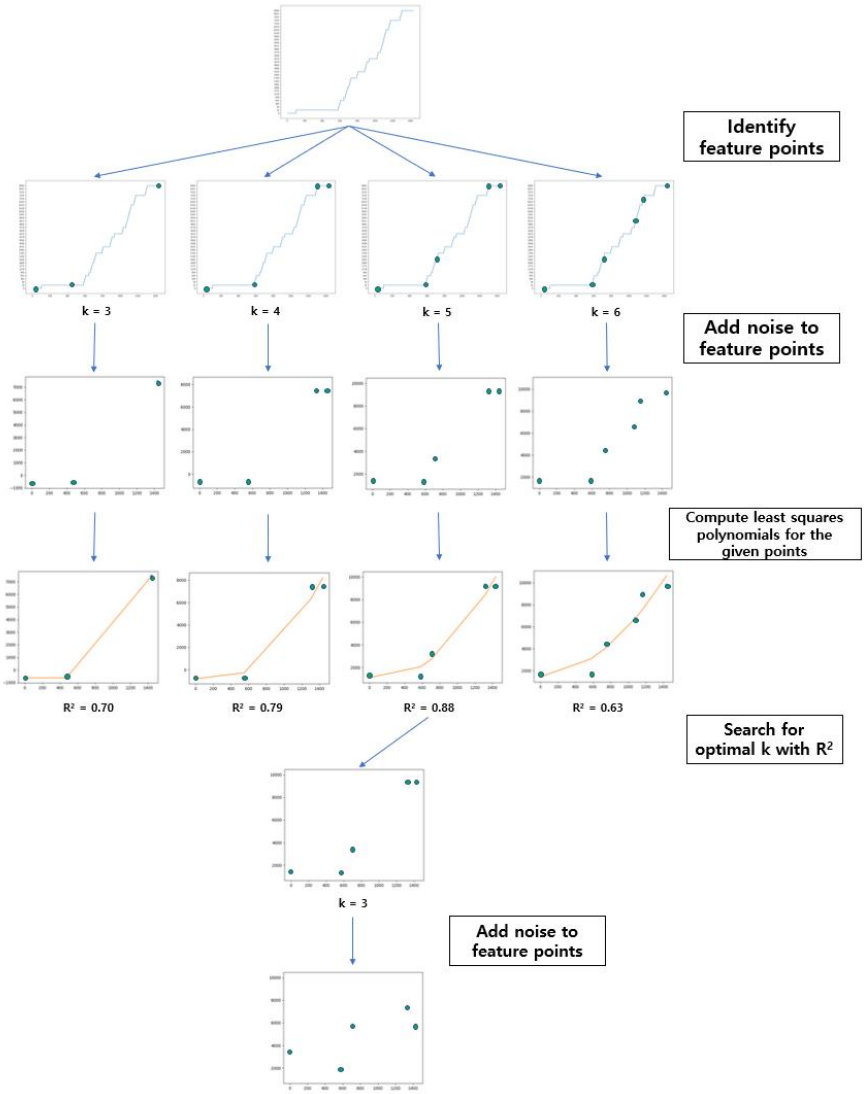


Fig. 1. Owner-side process of proposed approach

C'_{list}를 얻게 되고, C'_{list}를 바탕으로 2차원 회귀 모델을 만든다(lines 24 - 25). Fig. 3은 C_{list}와 도출된 2차원 회귀 모델의 R^2를 Fig. 2에 반환한다.

Fig. 2와 Fig. 3을 통해 u_i로부터 도출된 특징 점 데이터는 $bestC^i_{list} = \left(\left(t_{f_1}, x^i_{f_1}\right), \left(t_{f_2}, x^i_{f_2}\right), \cdots, \left(t_{f_{optimalk}}, x^i_{f_{optimalk}}\right)\right)$이다. 데이터 소유자는 $bestC^i_{list}$의 x^i_r에 지역 차분 프라이버시를 만족하도록 라플라스 매커니즘을 적용한 결과인 $pbestC^i_{list} = \left(\left(t_{f_1}, px^i_{f_1}\right), \left(t_{f_2}, px^i_{f_2}\right), \cdots, \left(t_{f_{optimalk}}, px^i_{f_{optimalk}}\right)\right)$를 데이터 수집가에게 전송한다. 3.2의 pu_i는 길이가 n이기 때문에 노이즈 분포가 $Lap\left(\frac{\Delta s}{\varepsilon/n}\right)$인 반면 $pbestC^i_{list}$는 길이가 $optimalk$인 데이터를 전송하게 되므로, $Lap\left(\frac{\Delta s}{\varepsilon/optimalk}\right)$에서 노이즈가 추출된다. 따라서 $pbestC^i_{list}$를 전송할 때 노이즈가 훨씬 적게 추가된다.

Algorithm 1 Finding optimal k

input: Range of k, Sequence of a health data s_i
output: Optimal k, Feature points C_{list}

1: $optimalk \leftarrow NULL$
2: $bestR^2 \leftarrow 0$
3: $bestC_{list} \leftarrow NULL$
4: **for** $i = 3 \ldots k$ **do**
5: $\quad tempC_{list}, tempR^2 \leftarrow SFP(i, s_i)$ ▷ Algorithm 2
6: $\quad$ **if** $bestR^2 < tempR^2$ **then**
7: $\quad\quad bestR^2 \leftarrow tempR^2$
8: $\quad\quad bestC_{list} \leftarrow tempC_{list}$
9: $\quad\quad optimalk \leftarrow i$
10: $\quad$ **end if**
11: **end for**
12: **return** $(optimalk, bestC_{list})$

Fig. 2. Pseudo-code for finding optimal k

Algorithm 2 Proposed Method for Searching Feature Points

input: The number of feature points k, Sequence of a health data s_i
output: Feature points C_{list}, R^2

// Step 1 : Identify feature points
$C_{list} \leftarrow NULL$
$P_{list} \leftarrow NULL$
$slp \leftarrow GetSecondtoLastPoint(k)$
$Endpoints_{list} \leftarrow GetEndPoints(k)$
$error_{min} \leftarrow \infty$
while $P[0] < Endpoints[0]$ **do**
  for $h \leftarrow P[slp]+1$ to $ListSize(s_i)$ **do**
    $curerror \leftarrow GetCurError(P_{list})$
    if $curerror < error_{min}$ **then**
      $error_{min} \leftarrow curerror$
      $UpdatePoints(C_{list}, P_{list})$
    end if
  end for
  $P_{slp} \leftarrow P_{slp}+1$
  for $h \leftarrow slp$ to 0 **do**
    if $P[h] = Endpoints[h]$ **then**
      $P[h-1] \leftarrow P[h-1]+1$
      $P[h] \leftarrow P[h-1]+1$
    end if
  end for
end while
// Step 2 : Compute the R^2
Generate random numbers $\sim$ Lap$(\frac{\Delta s}{\epsilon/k})$
r: sum of random numbers
$C'_{list} \leftarrow C_{list}+r$
Compute polynomial with C'_{list}
Compute R^2 between polynomial and original data
return (C_{list}, R^2)

Fig. 3. Pseudo-code for searching feature points and R^2 with step count data [16]

2. Data Aggregator-side Processing

데이터 수집가는 데이터 소유자로부터 변조된 특징 점 $pbestC^i_{list}$를 전송받는다. $pbestC^i_{list}$는 데이터 소유자 u_i의 최적의 특징 점 개수에 따라 길이가 서로 다르며, 데이터 수집가는 $AVG(px^i_r)$를 계산하기 위해, $pbestC^i_{list}$의 각 점 $\left(t_{f_m}, px^i_{f_m}\right), \left(t_{f_{m+1}}, px^i_{f_{m+1}}\right)$을 직선으로 연결한다 (Fig. 4).

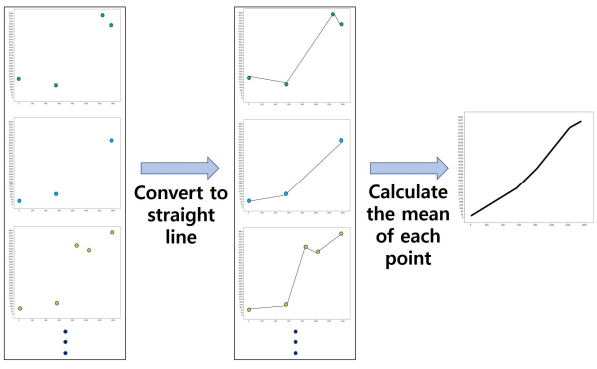


Fig. 4. Aggregator-side process of proposed approach

직선으로 연결하면 길이 n의 데이터 집합 pu_i를 계산할 수 있다. 데이터 수집가는 변조된 데이터 소유자 집단 PL로부터 수집한 px^i_r의 시간대별 평균 $AVG(px^i_r)$을 얻게 된다. 이처럼 데이터 수집가는 데이터 소유자로부터 변조된 특징 점 집단 pu_i를 전송받게 되므로, 데이터 유출에 대한 위험성 없이 민감한 건강 데이터를 수집할 수 있다.

V. Experiments and Results

1. Experiments

본 논문은 건강 데이터의 유용성을 확보하면서 사생활 침해 없이 수집하는 방법을 제안했다. 제안한 방법에 대한 유의성을 확인해 보기 위해 290명의 스마트밴드로부터 10분 단위로 10시부터 21시까지 수집한 누적 걸음 수 데이터로 실험을 진행했다. 데이터 사이즈에 따른 성능 평가를 위해 10, 100배 크기로 복제하여 사용하였으며, 다양한 ϵ 값에 따른 오차율 측정을 위해 ϵ을 0.5,1.0,2.0으로 설정하였다. 전역 민감도 Δs는 누적 걸음수의 최댓값 10,000과 최솟값 2,000의 차인 8,000을 사용했다. 데이터 수집가 측면에서 계산한 $AVG(px^i_r)$와 원본 통계치 $AVG(x^i_r)$를 비교하여 평균 절대 오차로 성능 평가를 진행하였고 평균 절대 오차 e의 수식은 다음과 같다. 실험에서는 n시간 측정한 n개의 데이터를 전송한 경우와 k개의 특징 점을 전송한 경우의 e를 비교하였다.

$$e = \frac{1}{n} \times \sum_{d=1}^{n} \left| AVG(x^i_r) - AVG(px^i_r) \right|$$

Table 1. An example of average error for proposed approach

data size \ ε	0.5	1.0	2.0
290×10^1	238460.91	120437.16	63884.69
290×10^2	7708.52	3872.07	1826.34

(a) naive method

data size \ ε	0.5	1.0	2.0
290×10^1	471.28	210.78	143.63
290×10^2	17.20	9.19	5.57

(b) optimal k

Table. 1에 따르면 n개의 데이터를 전송한 (a)보다 제안하는 특징 점 개수 알고리즘을 사용하여 최적의 특징 점을 찾아 전송한 (b)의 e가 낮은 것을 확인할 수 있다. 즉 모든 데이터가 아닌 특징 점을 추출하여 전송하는 방식이 데이터 유용성을 확보할 수 있음을 의미한다.

2. Results

Fig. 5는 특징 점 k를 고정 시킨 경우와 제안하는 알고리즘을 통해 최적의 특징 점 k를 전송받아 $AVG(x'_i)$를 계산한 통계치의 원본을 비교한 그래프이다. 제안하는 특징 점 추출 기법으로 최적의 특징 점 k개를 전송했을 때 원본 그래프와 가장 유사 하다는 것을 확인할 수 있다.

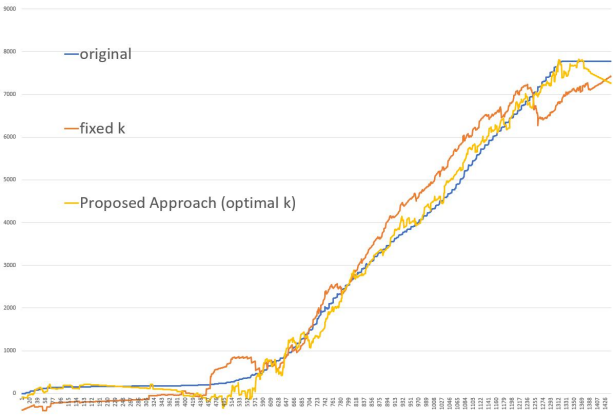
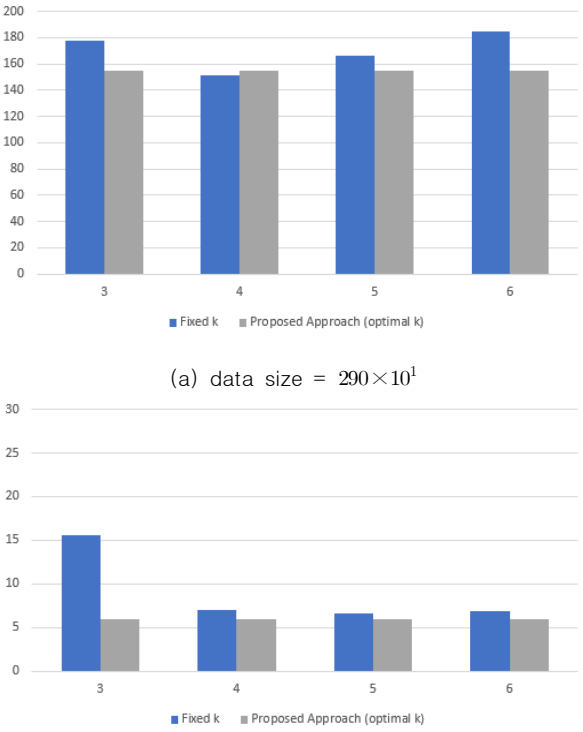


Fig. 5. An example of average error comparison between fixed k and optimal k (data size = 290×10^1, $\varepsilon=2.0$)

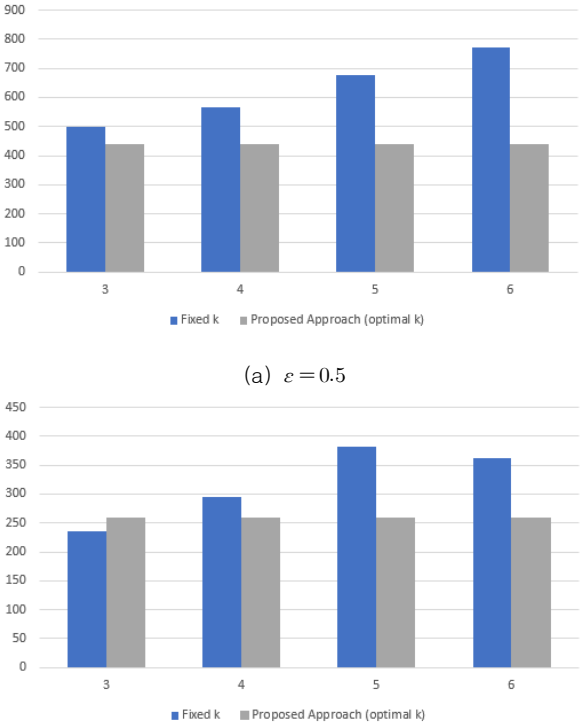
차분 프라이버시의 정의에 따라 데이터 사이즈가 클수록, 프라이버시 강도를 결정하는 ϵ값이 높을수록 원본 데이터와 유사하게 측정 되었다. Fig. 5는 데이터 사이즈에 따른 e의 차이를 나타낸다. 차분 프라이버시에서 오차율은 데이터 사이즈와 반비례 관계다. 이에 따라 ϵ을 2로 고정할 때, (a)보다 데이터 사이즈가 10배 더 큰 (b)가 더 적은 e를 보인다. Fig. 6은 데이터 사이즈를 290×10^1로 고정할 때 ϵ값에 따른 e의 변화를 보여준다. 차분 프라이버시에서 ϵ값이 작아질수록 추가되는 노이즈가 커지므로 프라이버시 보존 정도는 높아진다. 따라서 Fig. 6에서 ϵ값이 클수록 e가 줄어드는 것을 확인할 수 있다. ϵ값이 가장 작은 (a)의 e가 가장 높으며 ϵ값이 가장 큰 (c)의 e가 가장 낮다. 이처럼 프라이버시 강도와 데이터 유용성은 트레이드 오프 관계를 보이므로 적절한 ϵ값 설정이 중요하다. 마지막으로 Fig. 5와 Fig. 6 모두 k를 고정시킨 경우보다 제안하는 특징 점 개수 탐색 알고리즘으로 최적의 k를 적용한 경우가 더 낮은 e를 보였다. 이는 본 연구에서 제안하는 특징 점 추출 및 탐색 알고리즘을 통해 단조 증가 형태의 건강 데이터를 데이터 유용성을 확보하면서 개인 정보 침해 없이 수집할 수 있음을 의미한다.



(a) data size = 290×10^1

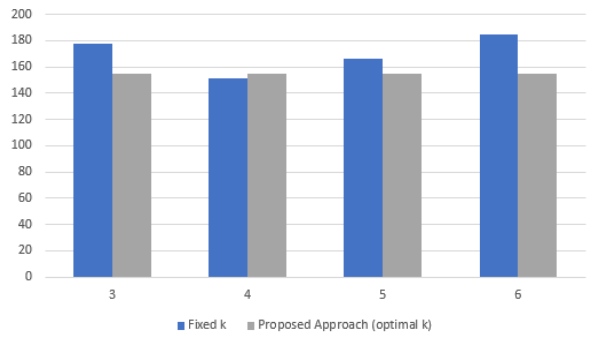
(b) data size = 290×10^2

Fig. 5. Average error comparison between fixed k and optimal k for different data size ($\varepsilon=2.0$)



(a) $\varepsilon=0.5$

(b) $\varepsilon=1.0$



(c) $\varepsilon = 2.0$

Fig. 6. Average error comparison between fixed k and optimal k for various privacy budget (data size = 290×10^1)

VI. Conclusions and Future Work

웨어러블 기기가 발전하면서 건강 상태를 시공간의 제약 없이 저비용으로 진단할 수 있게 되었다. 빅데이터와 각종 머신 러닝 기법들을 통해 심전도, 산소 포화도 등은 환자의 이상 신호에 빠르게 대처할 수 있도록 활용되고 있다. 이처럼 건강 데이터는 활용도가 높지만 공격자에게 오·남용될 경우 심각한 사생활 침해로 발전할 수 있다. 따라서 건강 데이터의 유용성은 확보하면서 개인 정보 침해를 방지하기 위한 기법이 필요하다. 본 연구에서는 지역 차분 프라이버시를 통해 데이터 소유자의 개인 정보를 보호하고, 특징 점 개수, 위치 탐색 알고리즘을 통해 데이터 유용성을 확보하는 방법에 대해 제안하였다. 선행 연구에서 추출한 특징 점 위치를 기반으로 하여 특징 점 개수 k를 탐색하면, 보다 정확한 통계치 확보가 가능하다. 본 연구에서는 걸음 수 데이터로 실험하여 건강 데이터에 적합한 최적의 특징 점 개수 k를 탐색한 후 데이터 수집자에게 전송하는 방법이 고정된 k를 전송하는 방법보다 오차율이 낮다는 것을 확인하였다.

REFERENCES

[1] Yang, Z., Zhou, Q., Lei, L., Zheng, K., & Xiang, W. (2016). An IoT-cloud based wearable ECG monitoring system for smart healthcare. Journal of medical systems, 40(12), 286.

[2] Hänsel, K., Wilde, N., Haddadi, H., & Alomainy, A. (2015, December). Challenges with current wearable technology in monitoring health data and providing positive behavioural support. In Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare (pp. 158-161).

[3] Chan, Y. F. Y., Bot, B. M., Zweig, M., Tignor, N., Ma, W., Suver, C., ... & Wang, P. (2018). The asthma mobile health study, smartphone data collected using ResearchKit. Scientific data, 5, 180096.

[4] Wang, S., Chang, X., Li, X., Long, G., Yao, L., & Sheng, Q. Z. (2016). Diagnosis code assignment using sparsity-based disease correlation embedding. IEEE Transactions on Knowledge and Data Engineering, 28(12), 3191-3202.

[5] Dimitrov, D. V. (2016). Medical internet of things and big data in healthcare. Healthcare informatics research, 22(3), 156-163.

[6] Wu, P. Y., Cheng, C. W., Kaddi, C. D., Venugopalan, J., Hoffman, R., & Wang, M. D. (2016). -Omic and electronic health record big data analytics for precision medicine. IEEE Transactions on Biomedical Engineering, 64(2), 263-273.

[7] Manogaran, G., & Lopez, D. (2018). Health data analytics using scalable logistic regression with stochastic gradient descent. International Journal of Advanced Intelligence Paradigms, 10(1-2), 118-132.

[8] Chaudhuri, S., Oudejans, D., Thompson, H. J., & Demiris, G. (2015). Real world accuracy and use of a wearable fall detection device by older adults. Journal of the American Geriatrics Society, 63(11), 2415.

[9] Takeda, R., Tadano, S., Todoh, M., Morikawa, M., Nakayasu, M., & Yoshinari, S. (2009). Gait analysis using gravitational acceleration measured by wearable sensors. Journal of biomechanics, 42(3), 223-233.

[10] Nath, N. D., Akhavian, R., & Behzadan, A. H. (2017). Ergonomic analysis of construction worker's body postures using wearable mobile sensors. Applied ergonomics, 62, 107-117.

[11] Beaulieu-Jones, B. K., Yuan, W., Finlayson, S. G., & Wu, Z. S. (2018). Privacy-preserving distributed deep learning for clinical data. arXiv preprint arXiv:1812.01484.

[12] Guan, Z., Lv, Z., Du, X., Wu, L., & Guizani, M. (2019). Achieving data utility-privacy tradeoff in Internet of medical things: A machine learning approach. Future Generation Computer Systems, 98, 60-68.

[13] Mohammed, N., Barouti, S., Alhadidi, D., & Chen, R. (2015, June). Secure and private management of healthcare databases for data mining. In 2015 IEEE 28th International Symposium on Computer-Based Medical Systems (pp. 191-196). IEEE.

[14] Tang, W., Ren, J., Deng, K., & Zhang, Y. (2019). Secure data aggregation of lightweight e-healthcare iot devices with fair incentives. IEEE Internet of Things Journal, 6(5), 8714-8726.

[15] Qin, Z., Yang, Y., Yu, T., Khalil, I., Xiao, X., & Ren, K. (2016, October). Heavy hitter estimation over set-valued data with local differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 192-203).

[16] Kim, J. W., Kim, D. H., & Jang, B. (2018). Application of local

differential privacy to collection of indoor positioning data. IEEE Access, 6, 4276-4286.

[17] McSherry, F. D. (2009, June). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 19-30).

[18] Kim, J. W., Jang, B., & Yoo, H. (2018). Privacy-preserving aggregation of personal health data streams. PloS one, 13(11).

[19] Moon, S. M., & Kim, J. W. (2020). Privacy-Preserving Method to Collect Health Data from Smartband. Journal of The Korea Society of Computer and Information, 25(4), 113-121.

Authors

Su-Mee Moon received the B.S. degree from Sangmyung University in 2019, where she is currently pursuing the master's degree with the Department of Computer Science. Her research mainly focuses on data privacy and Artificial Intelligence.

Jong-Wook Kim received the Ph.D. degree in Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group, Teradata, from 2010 to 2013. Dr. Kim is currently an Associate Professor with the Department of Computer Science at Sangmyung University. His primary research interests include the area of data privacy, distributed databases, and query optimization.