다차원 혈당 지표를 활용한 Inverted Pinnacle Skyline 방식 기반 당뇨병 위험군 환자 우선 선별

이준형[1], 조민서[1], 김종완[2*]
[1]삼육대학교 인공지능공학과 학부생
[2]삼육대학교 SW 융합교육원 교수

shymoncev@gmail.com, msfallsky@naver.com, kimj@syu.ac.kr

Priority Selection of High-Risk Diabetes Patients Based on the Inverted Pinnacle Skyline Method Using Multidimensional Glycemic Indicators

JunHyeong Lee[1], Min Seo Jo[1], Jongwan Kim[2]
[1]Dept. of Artificial Intelligence Convergence, Sahmyook University
[2]Software Convergence Education Center, Sahmyook University

요 약

본 논문에서는 다차원 혈당지표를 역전된 정점 스카이라인 질의(Inverted Pinnacle Skyline Query)에 적용하여 당뇨병 전증(前症)에 해당하는 환자 중 당뇨병 수치에 가장 근접한 환자를 식별하는 기법을 제안한다. 당뇨병 전증을 겪는 환자는 수치에 따른 적절한 조치가 취해지지 않으면 당뇨병으로 진행될 가능성이 높아진다. 환자의 치료에서 의료자원의 한정성과 환자의 위험도에 따른 우선순위 분류는 중요한 고려 요소이다. IPS 기법은 기존 Skyline 알고리즘의 지배관계를 반대로 정의하여, 당뇨병 전증 환자 중 가장 높은 수치를 지닌 즉, 당뇨병으로 진행될 가능성이 가장 높은 고위험군 데이터들을 식별을 목표로 한다. IPS 는 우선적 치료가 필요한 고위험군 환자들에게 당뇨병 수치 관리를 위한 조기개입을 가능하게 하며, 당뇨병 예방에 기여할 수 있을 것이다.

1. 서론

당뇨병 전증은 혈당 수치가 정상수치 보다 높지만 당뇨병으로 진단될 만큼 높지 않은 상태로, 당뇨병으로 발전할 가능성이 높은 단계이다. 당뇨병 전증의 환자들은 당뇨병에 기반한 심혈관 질환과 같은 합병증의 발생 위험이 높은 상태이며, 이는 조기 개입을 통한 당뇨병 예방이 중요하다.[1] 당뇨병 예방을 위한 조기 개입에서는 효율적인 의료 자원 배분을 위해 당뇨병 전증 환자 중 당뇨병 수치에 가장 근접한 인원을 식별하고, 해당 환자군에 대한 우선적인 치료 수단 제공이 필요시 된다.

본 논문에서는 역전된 정점 스카이라인 질의(Inverted Pinnacle Skyline, IPS) 기법을 사용하여 다차원 혈당 지표를 바탕으로 당뇨 전증 환자 중 당뇨병 수치에 가장 가까운 고위험군 환자들을 식별하는 방안을 제안한다. Skyline 알고리즘은 다차원 데이터에서 서로 지배관계에 속하지 않는 최적의 데이터를 식별한다. Skyline 에서 데이터 A 가 데이터 B 를 "지배한다" 고 판단하기 위해선 A 는 모든 차원에서 B 보다 작거나 같으며, 적어도 하나의 차원에서 더 작은 값을 가질 때, 데이터 A 가 B 를 지배한다고 판단한다. IPS 는 이러한 지배관계를 반대로 정의함으로써, 지배관계의 판단 기준을 크거나 같은 값을 기준으로 재정의한다. IPS 는 위험도가 가장 높은 데이터를 찾아내는 것을 목표로 하며, 본 연구에서 IPS 의 식별 목표는 당뇨병으로 진행될 가능성이 가장 높은 고위험군 데이터이다. 이를 통해 식별된 고위험군 환자에 대한 조기 개입으로 당뇨병 예방의 효과를 극대화할 수 있을 것으로 기대된다.

2. 당뇨병 전증 수치 범위

당뇨병을 진단하는 주요 혈당 수치는 공복혈당, 식후혈당, 당화혈색소이며, 각 수치는 당뇨병의 유무를 판단하는 중요한 기준이 된다. 3 가지의 수치 중 가장 높은 수치가 환자의 상태를 대표하게 되며, 당뇨병 전증 수치의 범위는 <표 1>과 같다.[2]

<표 1> 당뇨병 진단 주요 혈당 수치 범위

검사 유형	당뇨병 전증(Pre-diabetes)
공복 혈당 (Fasting Plasma Glucose, FPG)	100 ~ 125mg/dL
경구 포도당 (Oral Glucose Tolerance Test, OGTT)	140 ~ 199mg/dL
당화혈색소 (Glycated, Hemoglobin)	5.7% ~ 6.4%mg/dL

* 교신저자(Corresponding Author)

3. Inverted Pinnacle Skyline

스카이라인에서 두 데이터 포인트 p 와 r 이 주어진 데이터셋 Ds 에서 존재할 때, p 가 r 을 지배한다면, 이를 $p \prec r$ 로 나타낸다. 지배관계가 성립하기 위한 조건은 두가지이다. 첫째, 최소 하나의 차원 j 에서 p 가 r 보다 우월해야 한다. 둘째, 그 외의 모든 차원 i 에서 p 는 r 보다 같거나 더 작은 값을 가져야 한다. 이는 차원 j 를 제외한 모든 차원에서 p 가 r 보다 작거나 같은 값을 지님을 의미한다. 이를 정리하면 스카이라인의 지배관계는 수식(1)과 같다.

$$p \prec r \Leftrightarrow \exists j \in [1,d] \text{such that} p.dj < r.dj \text{and} \forall i \in [1,d] - \{j\}, p.di \leq r.di \quad (1)$$

IPS 에서의 지배관계는 위 스카이라인 지배관계의 성립 조건을 반대로 정의한다. P 와 r 의 지배의 정의는 $p \succ r$로 정의되며, 반전된 지배관계의 목표는 "가장 큰 값들로 구성된" 즉, 고위험군의 데이터로 집합된 스카이라인을 계산하기 위함이다. 이러한 IPS 의 수식은 아래 수식(2)과 같이 표현된다.

$$p \succ r \Leftrightarrow \exists j \in [1,d] \text{such that} p.dj > r.dj \text{and} \forall i \in [1,d] - \{j\}, p.di \geq r.di \quad (2)$$

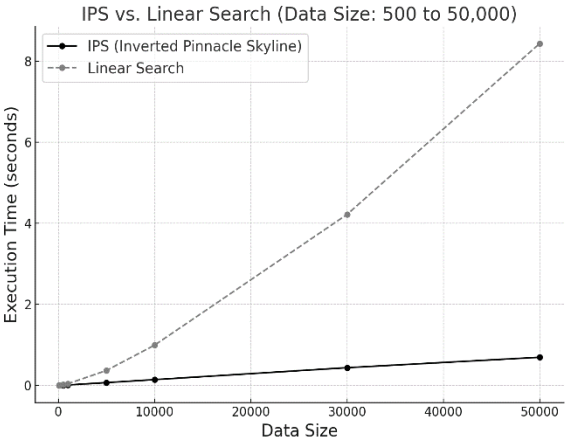
스카이라인의 지배관계인 수식(1)은 데이터 중 수치가 가장 낮은 데이터들을 선별하지만, IPS 의 지배관계인 수식(2)은 수식(1)의 지배관계를 반전함으로 수치가 가장 높은 데이터들을 선별한다. 이는 IPS 기법이 혈당 수치가 가장 높은 환자군을 식별하는 것을 의미한다. 식별된 데이터는 당뇨병 전증 환자 중 당뇨병으로 진행될 가능성이 가장 높은 환자들이다. IPS 의 알고리즘은 다음과 같다.

Algorithm: Inverted Pinnacle Skyline (IPS)

Input: Dataset D (FPG, OGTT, HbA1c)
Output: Set S containing IPS of high-risk prediabetes patients
1. Initialize S = ∅ (skyline set)
2. For each patient p ∈ D:
3. For each patient q ∈ S:
4. p dominates q: If p > q in at least one attribute (FPG, OGTT, or HbA1c) and p ≥ q in the others, remove q from S.
5. p is dominated by q: If p ≤ q in all attributes, skip to the next patient p.
6. If p is not dominated by any point in S, add p to S.
7. Return S.

4. 실험

실험에 사용된 소프트웨어는 Python3.12 이며, 데이터는 혈당 수치의 범위 중 정상 단계부터 당뇨병 전 단계까지의 수치를 기반하여 생성하였다. 비교 속성은 <표 1>의 검사유형을 속성으로 3 차원 비교를 진행하였으며, 500~50,000 개의 데이터를 사용하여 고위험군 환자 선별을 계산하였다. 성능 테스트는 환자군 식별을 위해 선형탐색과 IPS 를 비교하였다. 선형탐색은 전체 데이터를 순차적으로 단일 평가를 진행하고, 각 데이터는 독립적인 평가가 이루어진다. 두 기법의 구동 시간 비교 결과는 다음 그림(1)과 같다.



(그림 1) IPS 와 선형 탐색 구동 시간 비교

그림 1 에서는 데이터 수가 증가함에 따라 IPS 의 실행 시간이 선형 탐색 보다 더 빠르게 수행됨이 확인된다. 이는 IPS 의 연산이 불필요한 비교를 배제하여, 동일한 작업에서도 IPS 의 연산 속도가 더 빠르다는 것을 확인할 수 있다. 이는 IPS 가 대규모 데이터셋에서도 효율적이며 우수한 성능을 가짐을 의미한다.

5. 결론

스카이라인 질의는 범위를 기반한 데이터에 적합하지 않지만, 본 논문의 IPS 기법에서는 범위를 기반한 당뇨병 수치의 데이터에서도 위험 환자군 선별을 신속히 진행하며, 범위 기반 데이터 적용의 관점을 달리하였다. IPS 기법은 당뇨 전증의 고위험군 환자를 빠르게 식별함으로 신속한 진단과 예방에 기여할 수 있을 것이다.

감사의 글

본 연구는 2021 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업 지원을 받아 수행되었음 (2021-0-01440).

참고문헌

[1] Hostalek, U. (2019). Global epidemiology of prediabetes – present and future perspectives. *Clinical Diabetes and Endocrinology*, 5(1), 1-9.
[2] Hannon, T. S. (2020). Promoting prevention, identification, and treatment of prediabetes and type 2 diabetes in youth. *Pediatric Diabetes*, 21(2), 194-203.