한국컴퓨터정보학회 하계학술대회 논문집 제28권 제2호 (2020. 7)

메모리 효율적인 3군 매칭 알고리즘 구현

김동길[O], 정성재[*]
[O]분당서울대학교병원 빅데이터센터,
[*]분당서울대학교병원 빅데이터센터
e-mail: {r2403,sungjae.jung}@snubh.org[O*]

Memory Efficient Tri-Matching Algorithm

Donggil Kim[O], Sung Jae Jung[*]
[O]Big Data Center, Seoul National University Bundang Hospital,
[*]Big Data Center, Seoul National University Bundang Hospital

● 요 약 ●

세 군 매칭을 수행하여 관찰 데이터를 구축하고 통계분석에 기반한 연구를 수행하는 경우가 종종 발생한다. 매칭작업은 각 군에 속한 개체의 성향점수를 서로 비교해 거리가 가까운 짝을 찾아야 하므로 카테시안 곱 만큼의 경우의 수를 따져야 하는 문제이고, 메모리 소요가 크다. 특히 세 군 매칭은 세 쌍의 거리가 가까운 triplet을 찾는 문제로, 세 개체 사이에 존재하는 세 개의 거리를 따져야 하기 때문에 메모리 소요가 두 군 매칭에 비해 훨씬 크다. 각 군에 속한 개체가 늘어나면 메모리소요가 기하 급수적으로 늘어나게 된다.

R패키지에 포함된 TriMatch함수는 세 군 매칭 수행을 위해 가장 널리 사용되는 프로그램이다. 이 프로그램은 세 개체 사이의 세 개 거리가 가장 짧은 triplet을 찾는 방식으로 구현 되었다. 이 프로그램은 메모리 소요가 매우 커 각 군에 속한 개체의 수가 많아지면 메모리 부족 에러가 발생하는 경우가 많다. 본 연구에서는 세 군 매칭에 소요되는 메모리 소요를 줄일 수 있는 알고리즘을 제안하고자 한다. 이 알고리즘의 구현을 통해 각 군에 속한 개체가 늘어나도 안정적인 세 군 매칭 결과를 얻을 수 있을 것으로 기대한다.

키워드: 3군 매칭 (tri-matching), 성향점수 매칭(propensity score matching), 메모리 효율적 알고리즘(memory efficient algorithm)

I. Introduction

통계분석의 다양한 연구 틀 중, case-control 연구는 사례군(질병군)에 대한 대조군 매칭을 통하여 특정 위험요인에 대한 노출 빈도를 분석하는 연구이다[1]. 이때 로지스틱 회귀식을 통하여 개체별 성향점수를 구하고 성향점수가 비슷한 대조군을 찾아 위험요인 노출에 대한 관찰을 수행한다[2]. 경우에 따라서는 세 군 매칭을 수행하여 관찰 데이터를 구축하여 연구를 수행하기도 한다[3][4]. 매칭작업은 각 군에 속한 개체의 성향점수를 서로 비교해 거리가 가까운 짝을 찾아야 하므로 카테시안 곱 만큼의 경우의 수를 따져야 하는 문제이고, 메모리 소요가 크다. 특히 세 군 매칭은 세 쌍의 거리가 가까운 triplet을 찾는 문제로, 세 개체 사이에 존재하는 세 개의 거리를 따져야 하기 때문에 메모리 소요가 두 군 매칭에 비해 훨씬 크다. 각 군에 속한 개체가 늘어나면 메모리소요가 기하 급수적으로 늘어나게 된다.

R패키지에 포함된 TriMatch함수는 세 개체 사이의 세개 거리가 가장 짧은 triplet을 찾는 방식으로 구현된 프로그램이다. 이 프로그램을 이용하면 256 GB 메모리가 장착된 연산 시스템에서 각 군에 1000개의 개체가 포함된 크기의 데이터상에서 매칭연산이 가능하고 그 이상이 되면 메모리 부족 에러가 발생하는 것을 확인 하였다.

본 연구에서는 세 군 매칭에 소요되는 메모리 소요를 줄일 수 있는 알고리즘을 제안하고자 한다. 이 알고리즘의 구현을 통해 각 군에 속한 개체가 늘어나도 안정적인 3군 매칭 결과를 얻을 수 있을 것으로 기대한다.

II. Preliminaries

1. Related works

1.1 TriMatch R 패키지

R 패키지의 TriMatch알고리즘은 3군 매칭으로 가장 널리 사용되는 프로그램이다. 이 프로그램에서는 세 군으로부터 두 군을 고른 3가지 경우에 대한 각각의 회귀식을 만들고 각 개체에 대한 성향점수를 구하는 것으로 매칭 알고리즘의 동작을 시작 한다. 이 프로그램의 동작방식을 요약하면 다음과 같다.

1. 각 회귀식을 이용하여 각 군에 속한 개체의 성향점수를 구한다.
2. 개체간 거리 D1, D2, D3를 구한다.
3. 연결가능한 모든 경우의 수를 따져 개체간 거리의 합이 가장 작은 triplet을 구한다.
4. 표준화된(normalized) 개체간 거리가 0.25이상인 간선은 제외한다.

이 방식은 3단계 카테시안 곱 연산이 발생하여 메모리 소요가 매우 커지게 된다. 이것은 각 군에 속한 개체의 수가 많아지면 메모리 부족으로 결과를 얻지 못하는 경우가 많은 이유가 된다.

III. The Proposed Algorithm

본 연구에서는 메모리 효율적인 매칭 알고리즘을 제안하고자 한다. 이 알고리즘의 핵심동작은 다음과 같다.

1. TriMatch와 같은 방식으로 성향점수를 구한다.
2. Fig.1.의 1단계에 나타낸것과 같이 두군 매칭을 먼저 수행하여 최적 매칭(best matching) 간선의 집합을 구한다.
3. 구해놓은 각 간선별로 나머지 군에 속한 개체와 연결하여 삼각형의 합을 구한다(Fig.1. 의 2단계).
4. 삼각형의 둘레의 길이가 작은 순으로 triplet을 선정한다.

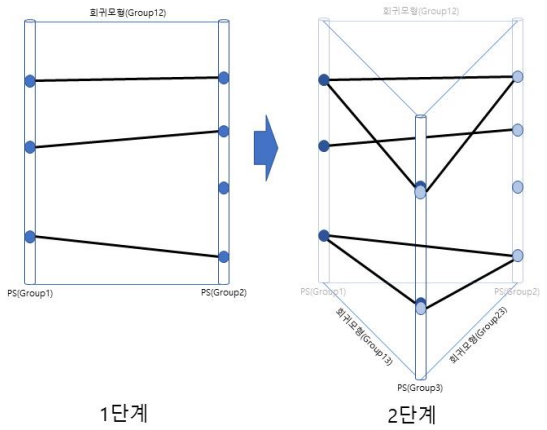

Fig. 1. 제안 Tri-Matching 알고리즘 동작

이 알고리즘은 R의 TriMatch 알고리즘에 비해 거리 비교단계가 한단계 줄어 메모리 효율을 크게 높일 수 있다. 이 알고리즘을 통하여 각 군에 속한 개체가 늘어나도 안정적으로 결과를 얻을 수 있는 프로그램을 구현할 수 있을 것으로 기대 한다.

IV. Conclusions

두 치료제의 효과를 비교하는 등의 사례 대조 연구에서는 3군 매칭을 통한 관찰 데이터 구축이 필요하다. 3군 매칭은 메모리 소요가 매우 크다. 이 연구에서는 메모리 효율적인 3군 매칭 알고리즘을 제안 하였다. 이 알고리즘이 프로그램으로 구현되면 안정된 3군 매칭 작업을 수행할 수 있고, 크기가 큰 3군 매칭 작업을 원활히 수행할 수 있을 것으로 기대 된다.

REFERENCES

[1] Aviva, Petrie and Caroline, Sabin "Medical Statistics at a Glance, 3rd Edition" November. 2013.

[2] Helmreich, J. E. and Pruzek, R. M. (2009). "PSAgra-phic s: An r package to support propensity score analysis." Journal of Statistical Software, 29(6):1–23.

[3] P. S. Nishtala. and D. Gnjidic. "'Real-world' haemorrhagic rates for warfarin and dabigatran using population-level data in New Zealand" International Journal of Cardiology, Vol. 203, pp.746-752, 15. January. 2016.

[4] H. Pierce and B. Al. Hussein. Al. Awamlh. "Impact of Body Mass Index on Outcomes Following Anato-mic GreenLight Laser Photoselective Vaporization of the Prostate" Journal of Endourology, Published On-line https://doi.org/10.1089/end.2020.0077 30. May. 2020.

[5] Bryer, J. M. "TriMatch: An R package for propensity score matching of non-binary treatments." 23. May. 2013.