국민건강보험 빅데이터 기반의 질병트렌드에 따른 지역 군집화 방법론 개발

국민건강보험 빅데이터 기반의 질병트렌드에 따른 지역 군집화 방법론 개발

이택호[1], 이수동[1], 전치혁*[1], 황명권[2], 이수철[3]

[1]포항공과대학교 산업경영공학과

[2]한국과학기술정보연구원

[3]목포대학교 경영학과

[1]{dlxorgh2, sudong_lee, chjun}@postech.ac.kr

[2]{mgh}@kisti.re.kr

[3]{suchullee}@mokpo.ac.kr

초 록

특정 지역의 질병 발생 추세는 지역의 산업, 사회, 또는 환경 등 여러 요인과 밀접한 관련이 있다. 특이한 질병 발생 추세를 야기하는 요인의 파악은 질병 예방 및 관리에 반드시 필요한 과정이다. 질병 특이트렌드에 따른 지역의 군집화는 질병의 특이트렌드를 야기하는 요인의 역추적에 기여할 수 있다. 본 연구에서는 급증/급감 특이치에 기반한 시계열 데이터 간 유사성 척도를 정의하고, 해당 척도를 이용한 시계열 군집화 절차를 제안한다. 제안하는 군집화 절차는 1) 질병 발생 급증/급감(특이치) 탐지, 2) 질병 발생의 특이치 기반 유사성 척도 계산, 그리고 3) 군집화 알고리즘 적용의 3단계로 이루어져 있다. 질병 트렌드 시계열 데이터를 구축 및 이용하여 위 절차에 따라 국내 지역들을 군집화함으로써, 질병 특이트렌드를 발생시키는 요인 추적의 기반을 마련한다.

1. 서론

질병 관리 및 예방을 위한 다양한 노력에도 불구하고, 질병 발생은 계속 늘어나고 있다. 또한, 다양한 질병의 발생 분포는 지역별로 다르게 나타나는데, 이는 지역의 환경 요인과 밀접한 관련이 있다. 지역의 환경 요인이 개인의 질병 발생에 영향력을 미치고 있지만, 환경 요인들은 개인에 의해 통제될 수 없는 사회적 특성이므로 포괄적인 관리가 요구된다 [9].

질병의 효율적 관리 및 예방을 위해 질병의 빈도와 분포에 영향을 미치는 환경 요인들을 규명하기 위한 노력은 오랫동안 계속되어 왔다. 대표적으로 영국의 존 스노우(John Snow)는 과거 콜레라의 유행 양상과 유럽 지역별 변화 양상을 비교하여, 특정 상수원들이 콜레라 유행의 근원임을 주장함으로써 오염된 물이나 식품 섭취를 통해 콜레라가 전염 된다는 사실을 밝혀냈다.

현대에 와서는 전염병 외에도 다양한 질병에 대해 환자 개인적 요소가 아닌 환경 요인을 찾는 것이 중요해지고 있다. 본 연구에서는 각 질병의 발생 횟수를 지역에 따라 월별로 정리함으로써, 질병과 지역마다 시계열 데이터를 구축한다. 각 질병트렌드 시계열 데이터 사이의 유사성을 기반으로 지역을 군집화하고, 군집 지역간 환경 요인의 차이를 분석함으로써 질병 발생의 빈도 및 분포의 차이를 야기하는 환경 요인들을 역추적하기 위한 기반을 마련한다.

2. 관련 연구

2.1. 시계열 군집화

군집화란 사전 지식 없이 속성이 유사한 데이터를 하나로 묶는 방법이다. 각 군집들은 군집 내 데이터 객체간 유사성을 최대화하고, 서로 다른 군집에 있는 객체와의 유사성은 최소화 하도록 구성된다. 이 때, 객체간 유사성을 측정하는 척도로는 일반적으로 비유사성을 측정하는 거리 척도를 대신 사용하므로, 각 군집들은 군집 내 데이터 객체간 거리는 최소화하고, 서로 다른 군집에 있는 객체와의 거리는 최대화하도록 구성된다.

시계열 데이터는 차원이 크고 간단히 다루기 어려우므로 군집화는 시계열 데이터로부터 유의미한 규칙을 발견하기 위한 방법으로 사용될 수 있다. 그러나, 시계열 데이터는 일반 데이터와 속성이 다르므로 일반적인 군집화 방법을 그대로 적용하기 어렵고, 이에 따라 다양한 분야에서의 시계열 군집화를 위한 방법들이 연구되어 왔다. 시계열 군집화는 일반적으로 1) 차원 축소 기법 2) 유사성 척도 계산, 그리고 3) 군집화 알고리즘 적용의 3단계로 이루어져 있다 [2].

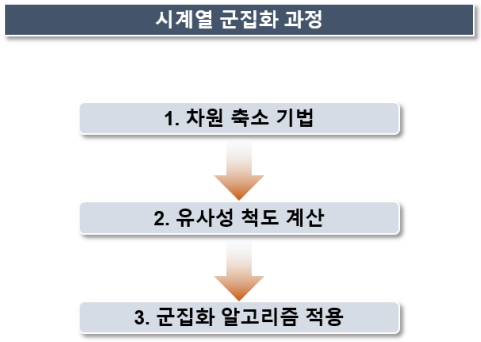

그림 1. 시계열 군집화 과정

2.2. 차원 축소(dimensionality reduction)

시계열 데이터를 낮은 차원의 공간에서 표현하도록 형태를 변환하는 것을 차원 축소라고 한다. 차원 축소 기법의 정의는 다음과 같다.

정의. (차원 축소 기법) 주어진 시계열 데이터 $T_i = \{t_1, \dots, t_N\}$에 대해, $n < N$을 만족하는 n에 대하여, $T_i = \{f_1, \dots, f_n\}$ 로 나타내는 방법을 차원 축소 기법이라고 한다. 또한, 원 공간에서 유사한 두 시계열 데이터는 변형된 공간에서도 유사하게 표현되어야 한다.

차원 축소 기법은 시계열 유사성 척도 계산에 중요한 역할을 하므로 대부분의 시계열 군집화 연구에서 사용되었다. 우선, 시계열 데이터는 대부분 데이터의 차원이 크기 때문에, 차원 축소 기법을 통해 계산 속도를 상당히 향상 시킬 수 있다. 또한, 시간 순서에 따라 시계열 데이터의 차원을 짝지어 거리를 계산하는 것은 데이터가 왜곡을 포함하고 있을 가능성을 고려하면, 직관적 해석 측면에서 단점을 가진다.

대표적인 차원 축소 기법으로는 다양한 분야에 걸쳐 사용되는 Singular Value Decomposition(SVD) 같은 matrix factorization 기법이 있다 [4]. 또한, 신호 처리 분야에서 널리 사용되는 Discrete Fourier Transform(DFT), Discrete Wavelet Transform(DWT)와 같은 기법들이 존재하고 [3], [8], 통계 분야의 Piecewise Aggregate Approximation(PAA), Adaptive Piecewise Constant Approximation(APCA) 등도 널리 사용된다 [7].

2.3. 시계열 유사성 척도

시계열 군집화 과정은 기본적으로 유사성 척도에 상당부분 의존한다. 유사성 척도란 두 객체 간 유사한 정도를 실수로 나타내는 함수를 의미한다. 유사성 척도에 대한 정확한 정의가 존재하기보다는 대부분의 경우 거리 함수의 역으로 정의되므로 사용할 거리 척도의 종류에 따라 군집화 결과가 결정된다.

2.3.1. 일반 거리 척도

시계열 데이터에 대해 일반 데이터 객체에 사용하는 거리 척도를 사용 할 수 있다. 대표적인 거리 척도는 유클리디안(Euclidean) 거리 척도이며, 각 시간 별 거리를 합하여 시계열 간 거리를 계산한다. 해당 척도는 단순하고 쉽게 계산이 가능하다는 장점이 있지만, 같은 길이를 가진 두 시계열에 대해서만 계산이 가능하고, 작은 오차에도 민감하게 변화한다.

$$d(X,Y) = \frac{1}{T}\sum_{t} X(t) - Y(t)$$

2.3.2. 동적 거리 척도

시간을 동적으로 연결하여 연결된 시간들에 대해 거리를 계산하는 척도를 사용 할 수 있다. 대표적인 방법으로는 동적 시간 워핑(Dynamic Time Warping, DTW)과 최장 공통 부분수열(Logest Common Sub-Sequence, LCSS)이 있다 [5], [6]. 해당 척

도는 서로 다른 길이의 시계열 간 거리 계산이 가능하며, 시간의 축에 따른 오차에 강건하다. 하지만, 일반 거리 척도에 비해 계산이 복잡하고, 데이터 진폭 오차가 클 경우 결과가 왜곡될 수 있다.

2.3.3. 차원 축소 기반 거리 척도

특정 차원 축소 기법과 결합하여 시계열 간 거리를 측정하는 척도를 사용할 수 있다. 해당 척도는 특정 차원 축소 기법을 상정한 상태에서 제안된 유사성 척도와 혹은 차원 축소 기법과 호환 가능한 기존의 거리 척도 모두 포함한다. 이 경우, 특정한 분야에 대한 맞춤형 거리 척도로 사용 가능하지만 수작업에 의한 적절한 설정이 반드시 수반되어야 한다. 해당 척도의 예로는 Threshold-based Queries(TQuEST)가 있다. 미리 정해둔 역치값을 이용하여 시계열 데이터를 역치값을 넘는 구간의 집합으로 간주하고, 구간의 집합이 유사할수록 시계열 간 거리가 작은 것으로 정의한다 [1]. 자세한 정의는 다음과 같다.

정의. (Threshold-Crossing Time Interval Sequence) 미리 정해둔 역치값 τ와 $l_j < t < u_j$인 모든 t에 대해 $x_t > \tau$를 만족시키는 (l_j, u_j)의 집합

정의. (Distance between Time Intervals) 두 구간 $\text{int}_1 = (l_1, u_1), \text{int}_2 = (l_2, u_2)$에 대해 두 구간 사이의 거리는 구간의 시작점과 끝점의 거리 $d_{(\text{int}_1, \text{int}_2)} = \sqrt{(l_1 - l_2)^2 + (u_1 - u_2)^2}$

정의. (Threshold-distance between two time series) 구간 집합 S_X와 S_Y로 나타낼 수 있는 두 시계열 데이터 X, Y에 대해 시계열 간 거리 $d_{TS}(X,Y) = \frac{1}{|S_X|}\sum_{s \in S_X} \min_{t \in S_Y} d(s,t) + \frac{1}{|S_Y|}\sum_{t \in S_Y} \min_{s \in S_X} d(t,s)$

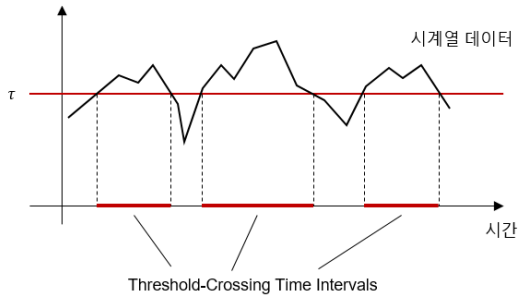

그림 2. TQuEST 유사성 척도

3. 제안 방법

3.1. 사용 데이터와 전처리

본 연구에서는 국민건강보험공단 12년(2002~2014) 코호트 데이터로부터 수집된 질병 발생 횟수를 이용해 지역별로 시계열 데이터를 구성하였다. 본 연구에서는 형성된 시계열 데이터로부터 증감 추세와 계절성을 제거하면 불규칙 오차만으로 구성된 시계열 데이터가 될 것으로 가정한다. 따라서, 증감 추세와 계절성을 제거한 시계열 데이터를 지수가중이동평균(Exponentially Weighted Moving Average, EWMA)방법으로 예측모형을 학습한다. 학습된 EWMA 모형과 실제 관측 값의 차이가 큰 경우(본 연구에서는 EWMA 표준오차 크기의 6배 이상으로 정의), 해당 시점에 특이치가 발생한 것으로 정의한다.

3.2. 특이치 기반 시계열 군집화

본 연구에서는 시계열 군집화 과정의 3단계를 이용하여 특이치 기반 시계열 군집화 과정을 제안한다. 차원 축소 기법으로는 특이치 탐사를 통해 각 시계열 데이터를 특이치 발생 시점의 집합으로 변형하고, TQuEST 기법의 거리 척도를 응용하여 특이치 발생 시점 간 거리의 평균으로 두 시계열 간 거리를 정의한다. 마지막으로, 특이치 기반 거리 계산 결과를 이용하여 일반적인 군집화 알고리즘을 적용하며, 본 연구에서는 군집화 알고리즘으로 K-medoids 알고리즘을 사용하였다.

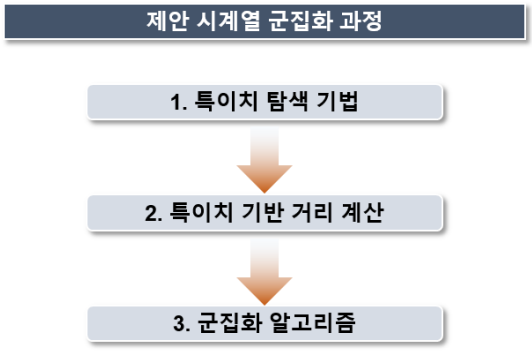

그림 3. 특이치 기반 시계열 군집화 과정

정의. (특이치 기반 거리) 특이치 집합 S_X와 S_Y로 나타낼 수 있는 두 시계열 데이터 X와 Y에 대해 특이치 기반 거리 $d(X,Y) =$

$$\frac{1}{|S_X|}\sum_{s\in S_X}\min_{t\in S_Y}\sqrt{(s-t)^2}+\frac{1}{|S_Y|}\sum_{t\in S_Y}\min_{s\in S_X}\sqrt{(t-s)^2}$$

4. 실험 결과

국민건강보험공단 코호트 데이터로부터 총 2,081개의 질병 중 질병코드 B, E, I, N의 총 317개 질병에 대해 지역별 군집화를 실시하였다. 지역은 행정구역분류코드 기준으로 시군구 단위에서 군집화 알고리즘을 적용하였으며, 데이터 신뢰성을 위해 질병별로 12년(144개월) 내에 발생한 환자의 수가 144명 이하인 지역은 알고리즘 적용에 앞서 제외하였다.

대표적으로 본태성원발성 고혈압(I10)의 경우, 전체 기간에 걸쳐 특이치가 고루 발생하는 군집과 일정 시기 이후 특이치가 갑자기 발생하는 군집으로 나뉘었다. 아래 그림에서 동그라미로 표시된 시점이 특이치에 해당한다. 아래 그림에서 분홍색에 해당하는 군집 1의 대표 지역은 '경상북도 안동시'로 전체 기간에 걸쳐 고루 특이치가 발생한다. 반면, 파란색으로 나타나는 군집 2의 대표 지역은 '충청남도 천안시'로 특정 시점(약 2007년 ~ 2008년)부터 특이치가 발생하기 시작한다.

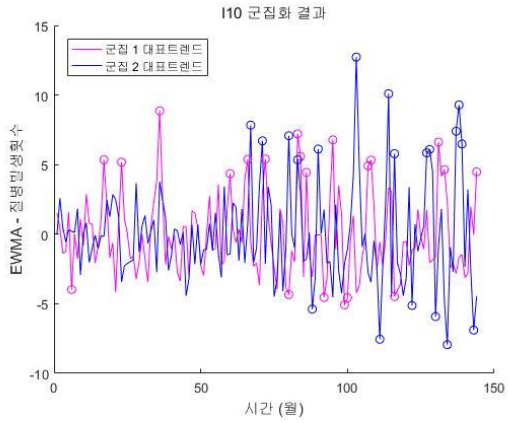

그림 4. 본태성원발성 고혈압(I10) 군집 결과

또한, 만성 신장병의 경우에도 본태성원발성 고혈압과 마찬가지로 전체 기간에 걸쳐 특이치가 고루 발생하는 군집과, 일정 시기 이후 특이치가 갑자기 발생하는 군집으로 나뉘었다. 아래 그림에서 분홍색에 해당하는 군집 2의 대표 지역은 '경상북도 군위군'으로 전체 기간에 걸쳐 고루 특이치가 발생한다. 반면, 파란색으로 나타나는 군집 1의 대표 지역은 '경기도 구리시'로 특정 시점(약 2007년 ~2008년)부터 특이치가 발생하기 시작한다.

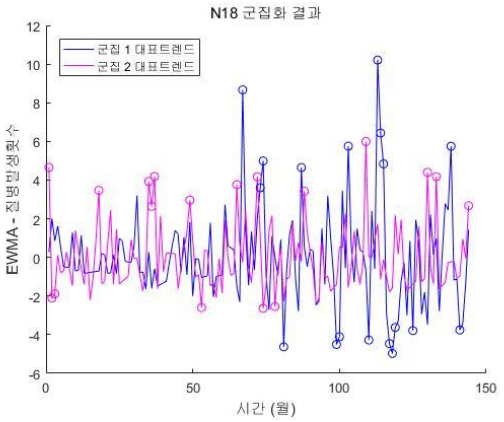

그림 5. 만성 신장병(N18) 군집 결과

5. 결론

본 연구에서는 국민건강보험공단 코호트 데이터를 이용하여 질병 및 지역별로 발생횟수를 시계열 데이터로 구성하였다. 특이치 탐지 기법을 기반으로 특이치 위치 기반 시계열 거리 척도를 정의하고, 이에 따라 특정 질병에 대해 서로 다른 트렌드를 보이는 지역끼리 군집화하였다. 각 질병에 대한 지역 군집화 결과를 이용하여 군집 내 지역들의 공통 요인들을 역추적함으로써 질병의 특이 발생을 야기하는 환경 요인 파악에 기여할 수 있다.

본 연구에서 사용한 시계열 군집화 알고리즘은 다음과 같은 한계를 가진다. 첫째, 제안하는 특이치 기반 거리 척도를 사용하기 위해서는 특이치 탐색이 선행되어야 하므로, 사용하는 특이치 탐색 알고리즘에 크게 의존한다. 둘째, 본 연구에서는 전체 12년 기간 동안 특이치 발생 위치가 유사할수록 두 질병트렌드가 유사하다고 가정하였다. 따라서, 탐지하고자 하는 질병트렌드의 정의를 달리할 경우에 대한 추가적인 연구가 필요하다. 마지막으로, 질병의 급증이나 급감 탐색이나 지역 군집에 대한 알려진 정보가 없으므로 결과에 대한 추가적인 검증이 필요하다.

6. 참고문헌

방의학회지, 38(3), 259-266.

[1] Aßfalg, J., Kriegel, H. P., Kröger, P., Kunath, P., Pryakhin, A., & Renz, M. (2006). Similarity search on time series based on threshold queries. *In International Conference on Extending Database Technology*, 276-294, Springer Berlin Heidelberg.

[2] Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–A decade review. *Information Systems, 53*, 16-38.

[3]Chen, X.Y., & Zhan, Y.Y. (2008) Multi-scale anomaly detection algorithm based on infrequent pattern of time series. *Journal of Computational and Applied Mathematics*, 214(1), 227-237.

[4] Klema, V., & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control, 25*(2), 164-176.

[5] Ratanamahatana, C. A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., & Das, G. (2010). Mining Time Series Data. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 1049-1077). Boston, MA: Springer US.

[6] Senin, P. (2008). Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 855*, 1-23.

[7] Wu, X. (1993). Adaptive split-and-merge segmentation based on piecewise least-square approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*(8), 808-815.

[8] Zhang, J., Tsui, F.C., Wagner, M.M., & Hogan, W.R. (2003) Detection of outbreaks from time series data using wavelet transform. *AMIA Annual Symposium Proceedings*, 2003, 748-752.

[9] 정성원, & 조영태. (2005). 한국적 특수성을 고려한 지역특성과 개인의 건강. 예