Journal of the Korean Data & Information Science Society
2016, **27**(3), 621–627

http://dx.doi.org/10.7465/jkdi.2016.27.3.621
한국데이터정보과학회지

범주형 시계열 자료의 군집화: 프로야구 자료의 사례 연구[†]

박노진[1]

[1]단국대학교 응용통계학과

접수 2016년 3월 12일, 수정 2016년 4월 27일, 게재확정 2016년 5월 11일

요 약

범주형 시계열 자료의 군집화에 대하여 정리해 보았다. 시계열 자료의 군집화는 일반적인 군집화에 시간을 고려해야하는 측면이 있다. 한편, 범주형 시계열 자료의 군집화에 대한 연구가 진행되었으나 현재 정리 요약된 국내외 논문을 찾기 어렵다. 본 논문에서는 범주형 시계열을 군집화 하는 몇 가지 방법들을 제시하고 그 방법들을 비교하기 위해 프로야구 데이터를 이용하였다. 프로야구 팀들 간에 어떤 팀이 특정 팀에 유독 약한 경기력을 보이는 경우가 있다. 국내 최강이라는 S팀이 유독 H팀에게 그런 경우가 그렇다. 2015년 S팀의 상대전적의 군집화를 통해 S팀과 H팀의 관계가 유별난 지를 밝히려 한다. 통계적으로 말하자면, 승/패로 이루어진 시계열 자료의 군집화를 수행하려는 것이다. 분석결과 S팀과 H팀과의 관계가 다른 팀들과의 관계에 비해 눈에 띄는 차이가 있음을 알 수 있었다.

주요용어: 범주형 시계열, 스펙트럼, 주기도, 주파수, 진화 나무.

1. 머리말

프로야구가 현재 국내에서 가장 인기 있는 스포츠로 자리 잡고 있다. 통계 전문가의 입장에서는 굉장히 재미있는 자료를 산출해내는 주요 데이터베이스라고 생각된다. 실제로 최근에도 Han 등 (2014)과 Lee (2015a)가 프로야구 데이터의 시계열 분석을 시도하였고 Cho와 Lee (2015), Kim과 Kim (2015) 그리고 Lee (2015b)는 다양한 고급 통계기법을 활용하여 프로야구 관련 자료의 분석을 시도하였다.

본 연구의 시작은 국내 프로야구 최강 구단이라는 S팀이 유독 H팀과의 전적에서 열세라는 점에서부터 시작되었다. 프로야구에서 소위 천적이라는 관계가 성립하는가에 대한 의문을 갖고 S팀과 상대팀들 간의 전적에 따른 군집화가 가능한가에 대해 분석을 시도하였다. S팀과 9개 상대팀과의 2015년도 전적을 이용하여 범주형 시계열의 군집화를 시도하려 하였다.

시계열 데이터 군집화와 관련하여 Lim 등 (2001), Park과 Kim (2008), Jung과 Jeon (2015)이 주로 경영 자료와 의학 자료에 대하여 시계열 군집화를 시도하여 좋은 결과를 얻었다. 시계열 자료의 군집화에 대한 자세한 요약은 Aghabozorgi 등 (2015)에서 확인할 수 있다. 시계열 데이터를 군집화하는 방법은 대개 세 가지로 이루어진다. (1) 자료의 시점에 따른 유사도 혹은 거리를 이용하는 분석 (2) 푸리에, 웨이블릿 변환을 통해 자료의 특징점을 찾아 이용한 분석 (3) 회귀, ARIMA, 마르코프, 신경망 같은 모델에 기반을 둔 분석으로 나눌 수 있다.

본 연구에서 다루고자 하는 자료가 연속형이 아닌 범주형이라는 점에서 위에 언급된 시계열 군집화 연구에 비해서 다른 방법이 요구된다. 본 연구에서는 기존의 여러 방법들에서 아이디어를 얻어 다음과 같은 세 가지 방법을 시도하였다.

[†] 이 논문은 2016년도 단국대학교 연구비에 의하여 수행되었음.
[1] (16890) 경기도 용인시 수지구 죽전로 152, 단국대학교, 교수. E-mail: rjpak@dankook.ac.kr

(1) 범주를 숫자로 코딩 변경한 후 군집화
(2) 유전자 서열 분석을 응용한 범주형 자료에 대한 군집화
(3) 범주형 주기도 (periodogram)를 이용한 군집화

2. 방법론

2.1. 범주의 숫자화에 의한 군집화

범주형 데이터의 군집방법으로 가장 원초적인 방법은 범주를 숫자로 코딩변경한 후 연속형 데이터로 인지하고 거리를 구하여 군집방법을 적용하는 것이다. 군집분석을 위해 일차적으로 비교 대상 데이터간 거리를 정의해야한다. 일반적으로 자료가 연속형일 때 군집분석에서는 유크리디안 거리가 대중적으로 사용되고 있다. 그런데 범주형 자료들 사이의 거리로 유크리디안 거리를 사용하는 것은 다소 정교하지 못한 면이 있다. 예컨대, A={1, 2, 3}, B={2, 3, 1} 그리고 C={1, 2, 6}이라는 3개의 데이터 집합이 있다고 하자. A와 B 그리고 A와 C의 유크리디안 거리는 각각 $\sqrt{6}$과 $\sqrt{9}$로 계산되어 A가 C보다 B에 가깝다고 할 수 있다. 하지만 숫자의 모양새를 고려할 때 A가 B보다 C에 유사하다고 볼 수도 있다.

따라서 범주형 데이터간의 거리, 예컨대 이분형 데이터간의 거리를 측정하는 것은 단순하게 생각할 수 없고 실제로 많은 연구자들에 의해 연구되어 왔다. Choi 등 (2010)의 서베이 논문에 의하면 범주형 자료에 대하여 총 76개의 유사성 (거리) 측도가 현재 개발되어 있는 것으로 되어있다. 그 측도들은 크게 '거리', '상관관계' 그리고 '비상관관계' 형태로 분류된다. 통계연구자들이 많이 사용하는 SPSS는 제곱 유크리디안 거리, Yule의 Q 등 27개의 측도를 탑재하고 있다. 이러한 측도들은 연속형으로 계산되며 연속형 자료에서 사용하는 군집방법을 사용하여 범주형 데이터의 군집분석을 수행한다.

예를 들어, $A = \{0, 1, 0, 1\}$이고 $B = \{1, 0, 0, 1\}$이라는 두 개의 데이터 집합의 제곱 유크리디안 거리는 $\sqrt{(b+c)^2} = 2$이고 Yule의 Q는 $2bc/(ad+bc) = 2/(1+1) = 1$가 된다 (Table 2.1).

Table 2.1 Example of binary table

		B	
		0	1
A	0	1 (a)	1 (b)
	1	1 (c)	1 (d)

2.2. 유전자 서열 분석 기법을 활용한 군집화

유전자 서열의 유사성을 연구하는 학자들에 의해 2.1에 언급된 방법과는 다소 다른 측면에서 범주형 데이터의 군집방법이 제안되었다. 여러 개의 서열들의 관계를 나무 구조로 표현하는 점에서 기본적으로 통계학에서 군집분석을 통해 개체간의 관계를 나무 형태로 표현하는 것과 유사하다. 하지만, 통계학의 군집분석 방법 보다 유전자 서열의 대체나 소실과 같은 순서상 또는 시간상 다소 고려해야할 내용들이 있다. 예컨대, Jukes와 Cantor (1969)의 아이디어를 소개하면 두 개의 서열 x와 y의 거리는 유크리디안 거리와는 다르게

$$d_{xy} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right), \ D\text{는 상이한 문자 (염기)의 비율}$$

로 정의된다.

거리를 계산한 후 나무를 만드는데, 나무는 뿌리를 갖는 나무와 뿌리를 갖지 않는 나무로 나눌 수 있다. 뿌리를 갖는 나무는 통계학적 용어로는 계층적 군집 분석에 의한 나무와 유사하고 뿌리를 갖지 않는

나무는 사회연결망 분석에서 사용하는 네트워크의 형태를 갖는다. 나무를 구성할 때 주로 사용하는 방법은 이웃연결방법 (neighboring joint algorithm)이다. 이 방법은 하단에서 상단으로 군집화하면서 진화나무를 만들어 가는 방법으로 Saitou와 Nei (1987)의 아이디어에 기반하고 있다. 이웃연결 알고리즘은 균형 최소 진화 기준에서 가장 최적화된 나무를 구성하는 알고리즘으로 흔히 탐욕적 알고리즘에 속한다. 거리행렬로 부터 특정한 방법으로 가지들을 다양한 조합으로 묶으면서 나무의 사이즈 혹은 길이를 구하되 사이즈 혹은 길이가 최소가 되도록 나무를 구성한다. 최종적으로 선택된 나무의 신뢰성을 측정하기 위해 부트스트랩방법을 사용하는데, 주어진 서열에서 일정한 크기의 반복이 허락된 랜덤표본을 선택하여 나무를 구성하는 작업을 수차례 수행하고 나무의 가지들이 동일한 형태를 유지하는 비율과 모비율에 대한 신뢰구간을 계산하여 모비율이 70%이상으로 추정될 때 나무 구조가 신뢰성이 있다고 판단한다 (Hillis와 Bull, 1993; Hillis 등, 1994). 본 논문에서는 이를 구현하기 위해 'ape'라는 R-package에서 'unrootedNJtree'와 'rootedNHtree' 함수를 사용하겠다.

2.3. 주파수 영역으로의 변환을 통한 군집화

때때로 시간에 따라 관찰된 데이터를 주파수 영역에서 분석하는 것이 유용할 때도 있다. 현재 가지고 있는 데이터가 시간영역에서는 범주형일지라도 이를 주파수영역으로 변환시키면 연속형 값을 갖는 데이터로 만들 수 있다. 시계열 혹은 신호가 등간격으로 N개의 시점에서 관측되어 시계열 $X = \{x_0, \ldots, x_{N-1}\}$가 주어졌을 때 주파수에 따른 데이터의 스펙트럼 밀도함수를 구할 수 있다 (Proakis와 Manolakis, 2006). 마치 확률밀도함수를 확률밀도함수 추정량으로 추정하듯이 스펙트럼 밀도함수도 실제적으로는 추정량인 주기도 (periodogram),

$$P(f) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x_n e^{-j2\pi nf}\right|^2, \quad -\frac{1}{2} < f \le \frac{1}{2}, \tag{2.1}$$

를 구하여 주요 주파수에 대한 강도 (amplitude)를 추정한다. ($\omega = 2\pi f$로 정의하여 $P(\omega)$로 쓰기도 한다.)

우리가 관심을 갖고 있는 범주형 시계열 데이터 집합들을 각각 주파수 영역으로 옮겨 연속형으로 만들고 군집화를 시도하려한다. 시계열 X에 대하여 어떤 시계열 $Y = \{y_{0,\ldots,}y_{N-1}\}$를 주파수 영역을 옮긴 후 두 시계열의 거리를 각각의 주기도를 이용하여

$$d(X,Y) = \left[\sum_{j=0}^{N-1}\{P_X(\omega_j) - P_Y(\omega_j)\}^2\right]^{1/2}, \ \omega_j = j2\pi/N \tag{2.2}$$

로 정의하고 이 거리를 기반으로 군집분석을 수행한다.

3. 데이터 분석

S팀에 대한 다른 9개 팀과의 2015년 정규시즌 승 (1)/패 (0) 자료를 사용하였다 (Table 3.1). 승/패 데이터를 연속형으로 생각하여 일반적인 군집분석을 수행하고 범주형으로 생각하여 유전자 진화 모형분석과 주파수 영역 분석을 수행하였다.

Figure 3.1에 연속형의 경우는 유크리디안 거리를 계산하고 범주형으로는 제곱 유크리디안 거리를 계산한 후, 완전연결법 (complete linkage)을 통해 군집화한 결과를 그려 넣었다. 그 결과 {D, H}, {SK, LG}, {NC, KT} 그리고 {{NX, KI}, LO}로 묶임을 확인할 수 있었다. S팀에 대한 D팀과 H팀의 전적이 유사하다고 하겠다. 본 논문에는 지면상 싣지 않았으나 실제로 거리의 종류와 연결법에 따라 군집의 결과가 다소 다르게 나타나는 것이 이 방법의 흠이라고 하겠다.

Table 3.1 Win/Loss of S team against the other teams

round	NC	D	NX	SK	H	KI	LO	LG	KT
1	1	1	0	1	0	1	1	1	1
2	1	1	1	0	1	1	1	0	1
3	1	1	1	0	0	0	1	0	1
4	0	1	1	1	1	1	0	0	1
5	1	0	0	0	0	0	0	1	1
6	0	1	0	1	0	0	0	1	1
7	1	0	1	1	0	1	1	1	0
8	0	1	0	1	0	0	1	1	0
9	0	1	0	1	1	0	1	1	0
10	1	0	1	1	0	1	1	1	1
11	1	1	1	1	1	0	0	1	1
12	1	1	1	0	1	1	1	1	1
13	0	1	0	0	1	0	0	0	1
14	1	0	0	0	0	0	0	1	1
15	1	1	1	1	0	1	1	0	1
16	1	0	1	0	0	1	1	1	1
W/L	11/5	11/5	9/7	9/7	6/10	8/8	10/6	11/5	13/3

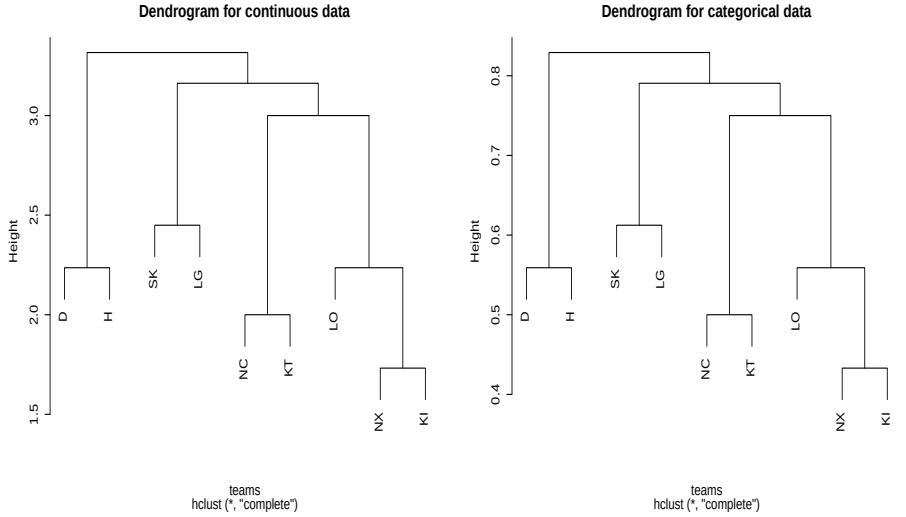

Figure 3.1 Dendrogram for continuous/categorical data

Figure 3.2에는 유전자 서열 분석 기법을 활용한 결과를 그려 넣었다. 팀 간의 진화적 위치를 명확히 보기 위해 S팀에 전승을 한 strong팀과 전패를 한 weak팀이라는 가상의 팀을 구성하여 같이 분석을 하였다. 그 결과 뿌리가 있는 경우와 없는 경우 모두 {strong, H}, {D>SK>LG}, {NC>LO>{KI, NX}}, {weak, KT}와 같이 군집화 할 수 있다. 이 경우 앞에서 시도한 군집분석에 비해 H팀이 다른 팀들과는 확실히 차별화됨을 볼 수 있다. H팀이 S팀에 전승했다고 가정한 strong팀과 가장 먼저 만나고 있다.

Figure 3.3에 각 팀의 전적에 대한 주기도와 그 들 사이의 거리를 이용한 군집분석 결과를 그려 넣었다. 주기도를 보면 LO팀이 가장 낮은 주파수를 갖고 있는데, 즉 S팀에 대하여 연승/연패의 주기가 길다는 것으로 실제로 2번의 3연패와 1번의 3연패를 기록한 것을 볼 수 있다. 반면 H팀은 큰 주파수를 갖고 있으며 실제로 S팀과 일진일퇴를 기록하고 있음을 볼 수 있다. 완전연결법을 사용한 군집결과는 {D,

H}, {NX, KI}, {LO, KT}, {NC, {SK, LG}}와 같다. 이는 Figure 3.1의 결과와 매우 유사하나 NC팀과 KT팀이 분리되어 있는데, 승/패 비율은 양 팀이 비슷하나 KT팀은 S팀에 대하여 6연패와 7연패의 기록을 갖고 있는데 이것이 주파수의 차이에 영향을 준 것으로 보인다. 주기도를 이용하는 경우 연패와 연승 같은 시간적 관계를 고려한 군집이 가능하다고 하겠다.

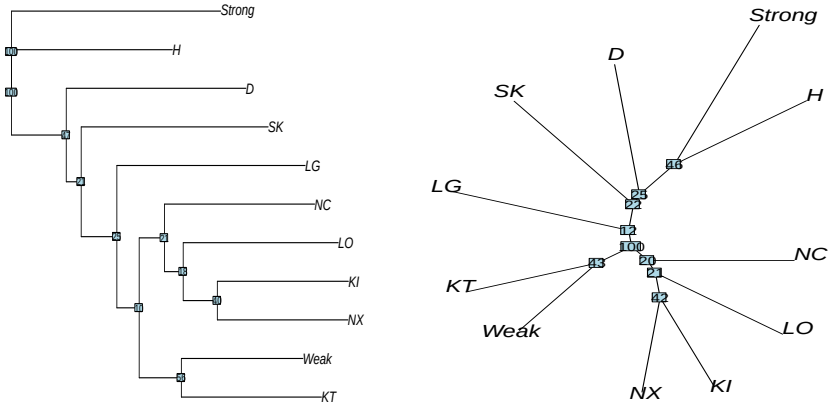

Figure 3.2 Rooted/ unrooted tree diagrams

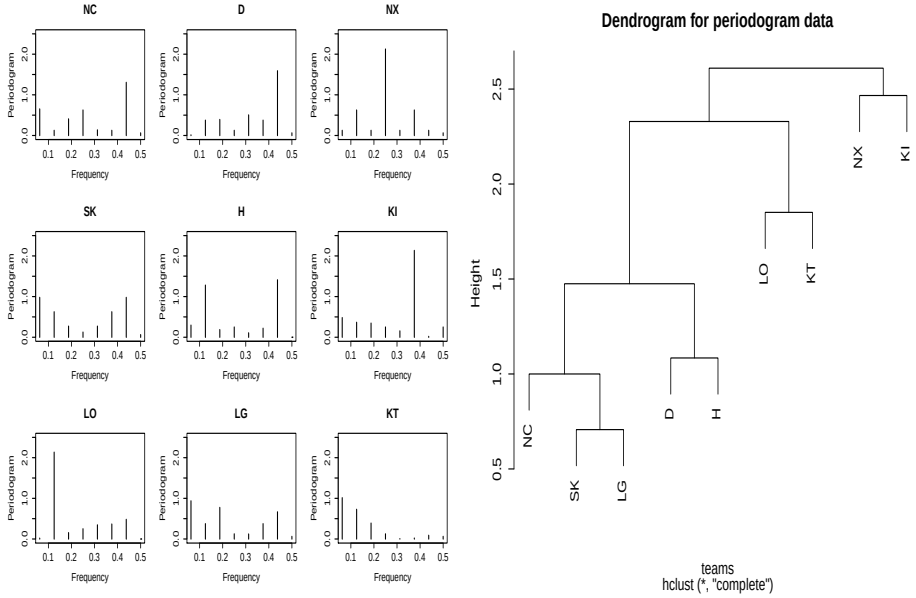

Figure 3.3 Periodograms and clustering on the frequency domain

4. 결론

본 논문에서는 범주형 시계열 군집화라는 다소 정리되지 않은 분야의 방법들을 제안하고 간단한 예제를 통해 그 활용성을 확인할 수 있었다. H팀에 대하여 S팀은 6승 10패로 다른 팀들에 비해 패배가 많

아 H팀을 S팀의 천적이라고도 한다. 범주형 시계열 군집화 분석 결과에 의해 H팀은 다른 팀들과 다른 승/패 데이터 구조를 갖고 있음을 확인할 수 있었다.

시간상의 흐름을 정확히 파악하려면 수평축을 시간으로 놓고 분석해야 할 것이나, 본 논문에서는 승패의 변화에 초점을 맞추어서 라운드를 수평축으로 놓고 낮은 수준에서 분석한 한계점이 있다. 또한, 승/패를 1/0으로 숫자화 하여 주기도를 구하는 경우 예를 들어 {000000111111}과 {111111000000}를 구분해 내지 못하는 주기도의 본질적 한계가 있다. 하지만 승/패의 변화 정도를 파악하고자 하는 좁은 범위의 분석에는 효과가 있다. 따라서 주기도외 다양한 군집분석을 통해 종합적으로 분석하는 것이 필요하겠다.

References

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

Cho, Y. J. and Lee, K. H. (2015). Bayesian estimation of the Korea professional baseball players' hitting ability based on the batting average. *Journal of the Korean Data & Information Science Society*, **26**, 197-207.

Choi, S. S., Cha, S. H. and Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Systems, Cybernetics and Informatics*, **8**, 43-48.

Han, G. H., Chung, J. and Yoo, J. K. (2014). A study on prediction for attendances of Korean probaseball games using covariates. *Journal of the Korean Data & Information Science Society*, **25**, 1481-1489.

Hillis, D. M. and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systems Biology*, **42**, 182-192.

Hillis, D. M., Huelsenbeck, J. P. and Cunningham, C. W. (1994). Application and accuracy of molecular phylogenesis. *Science*, **264**, 671-677.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of protein molecules in mammalian protein metabolism*, Academic Press, New York.

Jung, Y. A. and Jeon, J. H. (2015). A fusion of the period characterized and hierarchical bayesian techniques for efficient cluster analysis of time series data. *Journal of Digital Convergence*, **13**, 169-175.

Kim, N. K. and Kim, S. H. (2015). Comprehensive evaluation of baseball player's offensive ability by use of simulation. *Journal of the Korean Data & Information Science Society*, **26**, 865-874.

Lee, J. T. (2015a). Long term trends in the Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **26**, 1-10.

Lee, J. T. (2015b). Measuring the accuracy of the Pythagorean theorem in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **26**, 653-659.

Lim, J. Y., Zhang, B.-T. and Lee, K. M. (2001). Clustering fMRI time series using self-organizing map. *Proceeding of KFIS Fall Conference*, 251-254.

Park, M. S. and Kim, H. Y. (2008). Classification of precipitation data based on smoothed periodogram. *The Korean Journal of Applied Statistics*, **21**, 547-560.

Proakis, J. G. and Manolakis, D. K. (2006). *Digital signal processing: Principles, algorithms, and applications*, Prentice Hall, New York.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406-425.

Journal of the Korean Data & Information Science Society
2016, **27**(3), 621–627

http://dx.doi.org/10.7465/jkdi.2016.27.3.621
한국데이터정보과학회지

Categorical time series clustering: Case study of Korean pro-baseball data[†]

Ro Jin Pak[1]

[1]Department of Applied Statistics, Dankook University

Received 12 March 2016, revised 27 April 2016, accepted 11 May 2016

Abstract

A certain professional baseball team tends to be very weak against another particular team. For example, S team, the strongest team in Korea, is relatively weak to H team. In this paper, we carried out clustering the Korean baseball teams based on the records against the team S to investigate whether the pattern of the record of the team H is different from those of the other teams. The technique we have employed is 'time series clustering', or more specifically 'categorical time series clustering'. Three methods have been considered in this paper: (i) distance based method, (ii) genetic sequencing method and (iii) periodogram method. Each method has its own advantages and disadvantages to handle categorical time series, so that it is recommended to draw conclusion by considering the results from the above three methods altogether in a comprehensive manner.

Keywords: Categorical time series, evolutionary tree, frequency analysis, periodogram, spectral analysis.

[†] This research is supported by a research fund of Dankook University, 2016.
[1] Professor, Department of Applied Statistics, Dankook University, Yongin 16890, Korea.
E-mail: rjpak@dankook.ac.kr