The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.4.577
27(4), 577–588

Analyzing Financial Data from Banks and Savings Banks: Application of Bioinformatical Methods

Ro Jin Pak[a,1]

[a]Department of Applied Statistics, Dankook University

(Received April 17, 2014; Revised May 21, 2014; Accepted June 18, 2014)

Abstract

The collection and storage of a large volumes of data are becoming easier; however, the number of variables is sometimes more than the number of samples(objects). We now face the problem of dependency among variables(such as multicollinearity) due to the increased number of variables. We cannot apply various statistical methods without satisfying independency assumption. In order to overcome such a drawback we consider a categorizing (or discretizing) observations. We have a data set of financial indices from banks in Korea that contain 78 variables from 16 banks. Genetic sequence data is also a good example of a large data and there have been numerous statistical methods to handle it. We discover lots of useful bank information after we transform bank data into categorical data that resembles genetic sequence data and apply bioinformatic techniques.

Keywords: clustering, multicollinearity, phylogenetic tree, sequence alignment.

1. 서론

2008년 세계적 금융 위기 속에서 우리나라 저축은행들이 3차에 걸쳐 퇴출 되는 사건이 발생했고 많은 사람들이 저축은행의 문제점에 관심을 갖기 시작했다. 본 연구는 은행과 저축은행의 실질적 차이가 어디에 기인하는 지를 찾고자 하는 데서 시작되었다.

논문에서 분석하고자 하는 자료는 한국기업정보(KOCOinfo) 사이트에 등록된 은행 11개와 저축은행 5개의 재무지표 78개를 사용하였다. 은행과 저축은행과 관련된 중요 지표를 찾기 위해 SPSS를 이용하여 분석을 시도할 때 변수들 간의 지나친 중첩(혹은, 종속관계)으로 인해 분산-공분산 행렬이 계산 되지 않아 분석을 진행할 수 없다는 경고가 나타났다. 즉, 행렬의 행 또는 열이 선형 독립하지 않음으로 역행렬이 계산 되지 않는 등 계산의 어려움이 발생하게 되었다는 것이다. 선형 모형에서 이야기하는 심각한 다중 공선성이 존재한다는 의미가 되겠다.

최근 컴퓨터 및 여러 가지 측정 장비들로 인하여 대용량의 자료를 수집 및 저장하는 것이 가능하게 되면서 소위 빅데이터 분석, 데이터 마이닝을 통해 대용량 데이터를 분석하고 있다. 수많은 변수와 개체를 갖는 대용량 데이터를 분석하는 과정에서 통계적 추론에 기본이 되는 확률적 독립성이 보장 되지 않거나 무시해야 할 상황들이 발생한다.

The present research was conducted by the research fund of Dankook University in 2014.
[1]Department of Applied Statistics, Dankook University, Jukjun-Dong, Suji-Gu, Yongin 448-701, Korea.
E-mail: rjpak@dankook.ac.kr

Pak (2013)은 적절한 전처리 과정을 거쳐 이 문제를 해결할 수 있음을 보이기도 하였다. 한편, 본 논문에서는 대용량 데이터를 다루는 한 가지 방법으로 연속형 자료를 범주형 자료로 바꾸어 분석하면 위의 경고와 무관한 분석을 수행할 수 있지 않을까 하는 아이디어에서 시작되었다. 예컨대, 수열 A = $\{1, 2, 3\}$에 대하여 수열 B = $\{2, 3, 1\}$와 C = $\{1, 2, 6\}$를 숫자 순서대로 쌍으로 비교한다면 A와 B 그리고 A와 C의 유크리디안 거리는 각각 $\sqrt{6}$과 $\sqrt{9}$로 B가 C보다 A에 유사하다고 할 수 있다. 그러나 숫자들을 범주형으로 본다면 A와 C가 더 유사하다고 할 수도 있다. 데이터를 범주형으로 인식한 경우가 연속형으로 인식한 경우 보다 적절한 상황이 있을 수도 있다는 것이다.

연속형 자료를 범주형으로 변환하는 과정에서 정보의 손실이 다소 발생하지만 범주형 자료에만 가능한 분석 혹은 본격적인 분석에 앞선 탐색적 분석으로서의 효과가 분명히 있음을 본 연구를 통해 확인할 수 있었다. 특별히, 요즘 많은 관심을 받고 있는 범주형 자료인 유전자 서열을 분석하는 방법을 재무 지표 분석에 차용하여 의미 있는 결과를 발견할 수 있었다.

2. 방법론

앞서 생물정보학의 기법들을 활용하겠다고 했는데 그와 관련된 내용을 먼저 간단히 아래에 정리하겠다.

2.1. 서열정렬

생명정보학에서 서열정렬(sequence alignments)은 DNA, RNA 그리고 단백질이 갖는 서열 중에서 구조적, 기능적 그리고 진화적으로 유사한 서열을 찾아내는 과정을 의미한다. 그 과정은 서열을 행렬의 열벡터로 생각하여 수학적 혹은 통계적으로 구현된다. 다만, 수학에서의 행렬과 다른 점은 갭(gap)이라고 하는 비어 있는 셀(cell)을 고려해야 한다는 것이고 따라서 행렬을 다루는데 다소 복잡한 방법론이 필요하다는 것이다. 이러한 서열분석을 통해 서열 간의 유사성 혹은 상이성을 발견하고 생명 활동과의 관계를 파악하여 유전적 그리고 진화적 관계를 파악한다.

서열정렬은 분석 범위에 따라 지역정렬(local alignment)과 전역정렬(global alignment)로 나누고 분석 대상의 수에 따라 두 개의 서열을 대상으로 하는 쌍별 정렬(pairwise alignment)과 대상이 여러 개인 다중정렬(multiple alignement)로 나눌 수 있다.

예를 들어, 'AATCTATA'와 'AAGATA'라는 두 개의 핵산 서열이 있다고 한다면 이들 사이에 예컨대

$$\begin{array}{cc} AATCTATA & AATCTATA \\ AAGATA & AA\quad GATA \end{array}$$

와 같은 쌍별서열정렬이 가능한데 이 둘 중 어느 것이 더 의미가 있는지를 판단해야한다. 이를 위해 일치하는 경우, 불일치하는 경우 그리고 갭이 있는 경우에 따라 점수를 차등하여 정렬의 정도를 수치화한다. 만일 일치하면 1점, 불일치하면 0점 그리고 갭이 존재하면 -1을 부여한다면 첫 번째 정렬은 2점 그리고 두 번째 정렬은 3점이 된다. 따라서 이런 점수 방식이라면 두 번째 정렬이 보다 의미가 있다고 하겠다.

한편, 두 개의 서열 'FTFTALILLAVAV'와 'FTALLLAAV'에 대해 지역정렬과 전역정렬을 다음과 같은 정열이 가능하다.

$$\begin{array}{cc} \text{global alignment} & \text{local alignment} \\ FTFTALILLAVAV & FTFTALILL\ \ AVAV \\ F\quad TAL\quad LA\ \ AV & FTAL\ \ LLAAV \end{array}$$

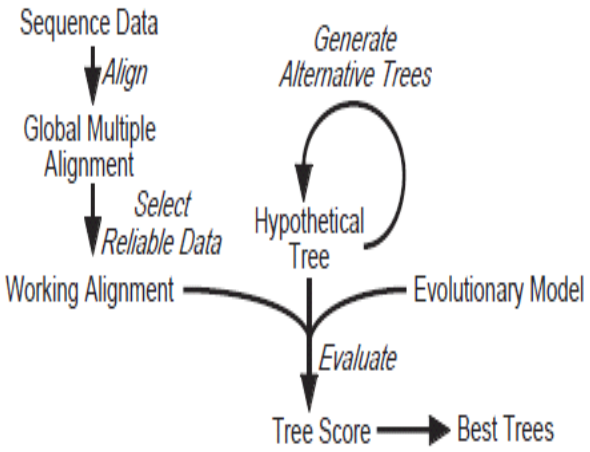


Figure 2.1. Construction of an evolutionary tree (Olsen, 2013).

전역정렬은 되도록이면 끝단을 갭으로 정렬하지 않으려 한다. 서열 중간 중간에 갭을 효과적으로 삽입함으로 모든 서열들의 처음과 끝을 맞추려 한다. 한편, 지역정렬은 가장 유사도가 높은 상태로 정렬을 시도한다. 전역정렬을 위해 Needleman-Wunsch 알고리즘이 지역정렬을 위해 Smith-Waterman 알고리즘이 주로 사용된다. 다중서열정렬은 쌍별서열정렬을 확장한 것으로 두 개 이상의 서열을 한 번에 정렬하고자 할 때 사용된다. 이것은 주로 많은 서열들이 공통적으로 갖는 부분을 확인하고자 할 때 사용된다. 물론 계산상으로 매우 복잡하고 어려운 작업이지만 수학의 행렬을 활용하는 등 다양한 방법으로 진화와 같은 매우 중요한 생물학적 발견을 위해 사용되고 있다 (Fitch, 1966). 알고리즘에 대한 자세한 설명은 Krane과 Raymer (2003)를 참조하기 바란다.

2.2. 진화 나무

생물정보학에서 나무를 만드는 과정은 아래 Figure 2.1과 같다. 먼저, 관심 있는 서열들을 지역 혹은 전역 정렬하여 중요한 서열을 찾거나 불필요한 서열을 제거하는 전처리 과정을 수행한 후 일차적으로 나무를 구성하고 이를 가설적 설정된 나무와 비교하는 과정을 거쳐 가장 적절한 나무를 찾는다 (Baldauf, 2003).

여러 개의 서열들의 관계를 나무 구조로 표현하는 것은 기본적으로 통계학에서 군집분석을 통해 개체간의 관계를 나무 형태로 표현하는 것과 유사하다. 다만, 통계학의 군집분석 결과 보다 다소 정교한 나무 형태를 구하기 위해 생물학적 아이디어에 기반을 두는 목적 추구형 혹은 모델 기반형 방법들을 사용한다 (Table 2.1).

군집분석을 위해 일차적으로 서열간 거리를 정의해야 하는데, 일반적인 군집분석에서는 유크리디안 거리가 대중적으로 사용되는데 이는 연속형 자료일 때 사용할 수 있는 방법이고, 두 개의 서열 사이의 거리를 측정하는 방법은 많은 연구자들 (Jukes와 Cantor, 1969; Kimura, 1980, 1981; Felsenstein, 1981; Barry와 Hartigan, 1987; Jin과 Nei, 1990; Tamura와 Nei, 1993)에 의해 시도 되었다. 초기의 방법인 Jukes와 Cantor (1969)의 아이디어를 소개하면 두 개의 서열 x와 y의 거리는 유크리디안 거리와는 다

Table 2.1. Algorithms on phylogenetic trees (Olsen, 2013).

Clustering	UPGMA(Unweighted Pair Group Method with Arithmetic Mean) WPGMA(Weighted Pair Group Method with Arithmetic Mean) neighbor-joining single-linkage, complete-linkage, average-linkage
Objective criterion-based (model based)	least-squares distance minimum evolution, maximum likelihood, maximum parsimony Bayesian

르게

$$d_{xy} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right)$$

(D는 상이한 문자(염기)의 비율)로 정의된다.

진화 나무는 뿌리를 갖는 나무(rooted tree)와 뿌리를 갖지 않는 나무(unrooted tree)로 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 이웃연결(neighbor joining)방법이다. 이 방법은 하단에서 상단으로 군집화하면서 진화나무를 만들어 가는 방법으로 Saitou와 Nei (1987)의 아이디어에 기반하고 있다. 이웃연결 알고리즘은 균형 최소 진화(balanced minimum evolution) 기준에서 가장 최적화된 나무를 구성하는 알고리즘이나 흔히 탐욕적(greedy) 알고리즘에 속한다. 거리행렬로부터 특정한 방법으로 가지들을 다양한 조합으로 묶으면서 나무의 사이즈 혹은 길이를 구하되 사이즈 혹은 길이가 최소가 되도록 나무를 구성한다. 이를 구현하기 위해 'ape'라는 R-package에서 'unrootedNJtree'와 'rootedNHtree' 함수를 사용할 수 있다.

최종적으로 선택된 나무의 신뢰성(reliability)을 측정하기 위해 부트스트랩(bootstrap)방법을 사용한다. 부트스트랩을 통해서는 나무 구조의 신뢰성 혹은 안정성(stability)을 확보하려는 것이다. 주어진 서열에서 일정한 크기의 반복이 허락 된 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 (Hillis 등, 1994).

3. 데이터 분석

한국기업정보(KOCOinfo)사이트에 등록된 은행 11개(b1, . . . , b11)와 저축은행 5개(sb1, . . . , sb5)의 재무지표 78개를 수집하였다. 78개 지표는 수익성, 생산성, 성장성, 활동성 그리고 안정성 모두 다섯 영역에 대하여 각각 35, 14, 13, 6과 10개의 지표들로 이루어져 있다 (Table 3.1). 본 논문에서는 저축은행 문제가 가장 심각했던 2010년도 자료를 사용하였다. 지표간의 상관관계는 영역 내 지표 간에 대체로 높고 수익성과 성장성 그리고 생산성과 안정성 내의 지표들 간에 상관계수가 높게 나왔다. 총 78개 지표들 중에서 6개 지표들(예수금평균이자율, 평균배당률, 부가가치, 2007년 이전 발생 종업원 1인당 경상이익, 종업원 1인당 대출채권, 종업원 1인당 인건비 증가율)을 제외하고는 분산팽창지수가 10보다 크게 계산되었다. 즉, 지표들 간에 심각한 공선성이 존재함을 알 수 있다. 따라서 본 논문에서는 앞서 제시한 범주화 방법을 사용하여 분석을 시도하고자 한다.

데이터 변환은 먼저 변수(지표)별로 16개 은행 자료를 표준화하고 25분위수 이하를 'T', 25분위수 초과 50분위수 이하를 'G', 50분위수 초과 75분위수 이하를 'C', 75분위수 초과를 'A'로 범주화하였다. 한편

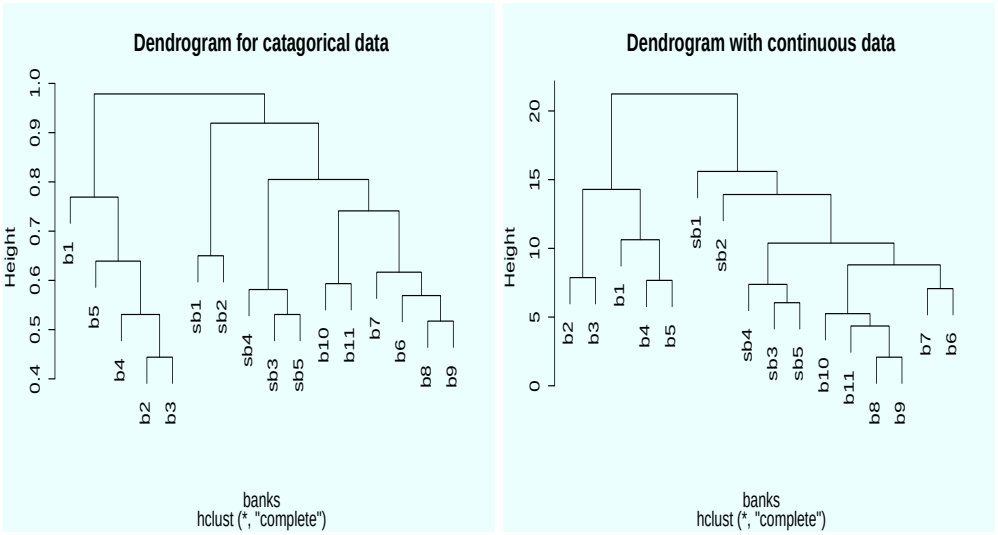
Table 3.1. List of variables under five subjects

profitability	productivity
1 Operating Revenues to Operating Expenses	1 Value Added
2 Operating Income to Operating Revenues	2 Value Added Per Employee
3 Ordinary Income to Operating Revenues (Prior to 2007)	3 Operating Revenues Per Employee
4 Net Income to Operating Revenues	4 Operating Income Per Employee
5 Total Income to Total Capital	5 Ordinary Income Per Employee (Prior to 2007)
6 Operating Income to Total Capital	6 Net Income Per Employee
7 Ordinary Income to Total Capital (Prior to 2007)	7 Personnel Expenses Per Employee
8 Net Income to Total Capital	8 Deposits Per Employee
9 Operating Income to Stockholder's Equity	9 Loans Per Employee
10 Ordinary Income to Stockholder's Equity (Prior to 2007)	10 Liabilities & Stockholder's Equity Per Employee
11 Net Income to Stockholder's Equity	11 Ratio of Value Added to Liabilities & Stockholder's Equity
12 Operating Income to Capital Stock	12 Ratio of Value Added to Net Sales
13 Ordinary Income to Capital Stock (Prior to 2007)	13 Ratio of Personnel Expenses to Value Added
14 Net Income to Capital Stock	14 1 Minus Personnel Expenses to Value Added
15 Operating Expenses to Operating Revenues	**growth**
16 Operating Expenses to Deposits	1 Total Capital Growth Rate
17 Non-Operating Income to Operating Revenues	2 Tangible Assets Growth Rate
18 Total Expenses to Total Revenue	3 Stockholder's Equity Growth Rate
19 Personnel Expenses to Total Expenses	4 Operating Income Growth Rate
20 Taxes & Dues to Total Expenses	5 Ordinary Income Growth Rate (Prior to 2007)
21 Accumulated Earning Ratio	6 Net Income Growth Rate
22 Accumulated Earning to Stockholder's Equity	7 Loans Growth Rate
23 Additional Paid-In Capital & Retained Earning to Stock holder's Equity	8 Deposits Growth Rate
	9 Operating Revenues Growth Rate
24 The Average Rate of Interest on Loans Receivable	10 Value Added Per Employee Growth Rate
25 Decrease in The Average Amount of Interest Accrued on Deposits Received	11 Employees Growth Rate
	12 Operating Revenues Growth Rate Per Employee
26 Dividends to Capital Stock	13 Personnel Expenses Growth Rate Per Employee
27 Dividends to Stockholder's Equity	**activity**
28 Dividends to Net Income	1 Total Assets Turnover
29 Operating Revenues Per Share	2 Stockholder's Equity Turnover
30 Earnings Per Share	3 Capital Stock Turnover
31 Ordinary Income Per Share (Prior to 2007)	4 Debt
32 Cash Flow Per Share	5 Loans Turnover
33 Book-value Per Share Turnover	6 Deposits Turnover
34 Reserves Ratio	**stability**
35 Operating Income Par Value	1 Tangible Assets to Total Assets
	2 Loans to Deposits
	3 Stockholder's Equity to Total Assets
	4 Debt to Total Assets
	5 Total Borrowings & Bonds Payable to Total Liabilities
	6 Total Borrowings & Bonds Payable to Total Assets
	7 Deposits to Stockholder's Equity
	8 Liabilities Ratio
	9 Reserves to Total Assets
	10 Cash Flows from Investing Activities Rates

모든 문자가 'A', 'C', 'G' 혹은 'T'로 각각 이루어진 가상의 은행 A, C, G 그리고 T를 구성하였다. 예컨대, 은행 A는 모든 지표들에서 최상의 범주값('A')을 갖는 가상 은행이라고 하겠다. 유전자 서열 형태로 범주화된 은행 자료에 대해 전역 정렬(global alignment), 지역 정렬(local alignment), 뿌리 없는 나무(unrooted tree) 그리고 뿌리 있는 나무(rooted tree)분석을 수행하였다.

3.1. 군집분석

범주화된 자료의 16개 은행 간 거리를 앞서 언급한 Jukes와 Cantor (1969)의 방법에 따라 거리를 구하고 최장거리법(complete linkage)에 의해 구한 나무(덴드로그램)와 연속형으로 이루어진 원자료들의 유크리디안 거리를 최장거리법으로 구한 나무을 그려보았다 (Figure 3.1). 두 그림을 보면 대략적으로 (b1, b2, b3, b4, b5), (sb1, sb2), (sb3, sb4, sb5) 그리고 (b6, b7, b8, b9, b10, b11)와 같이 4그룹으로 묶인다고 할 수 있겠다. 실제 첫 번째 그룹은 전국에 걸쳐 점포를 갖고 있는 거대은행들이 속해 있고 두 번째와 세 번째는 저축은행들 그리고 네 번째는 중소 규모의 은행들과 지방은행들로 이루어져 있다. 지면상 최장거리법의 결과만 수록했지만 연결법을 달리하여 분석을 해보았는데 조금씩 차이가 있기는 하



(a) Categorical (b) Continuous

Figure 3.1. Clustering with categorical data and continuous data

지만 범주형 자료와 연속형 자료를 사용한 결과가 매우 유사함을 관찰할 수 있었다.

3.2. 진화 나무

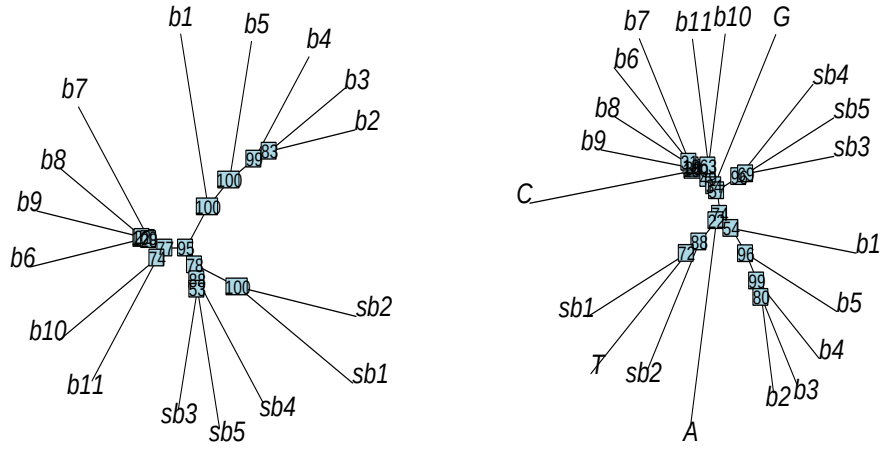
먼저 뿌리를 갖지 않는 나무 그림(Figure 3.2)을 보면 발생의 순서를 포함하여 {b1 > b5 > b4 > (b2, b3)}, {(b10, b11) > b7 > (b6, b8, b9)}, {(sb1, sb2) > sb4 > (sb3, sb5)}와 같이 세 집단으로 나뉘고 있다. 첫 번째 집단은 거대은행, 두 번째 집단은 중소은행 그리고 세 번째 집단은 저축은행들로 구성되어 있다. 가상 은행들과 함께 나무를 그려 보면, 첫 번째 집단은 A를 진화의 중심으로 하는 거대은행, 두 번째 집단은 C를 중심으로 하는 중소/지방은행, 세 번째 집단은 G를 따르는 저축은행(sb3, sb4, sb5) 그리고 마지막 집단은 T를 시점으로 저축은행(sb1, sb2)이 이어지는 방식으로 구성된다.

전체적으로 두 개의 나무가 유사한 결과를 보여주고 있다. 가상 은행 A가 매우 우수한 형질을 갖고 있다고 볼 때, A에 가장 근접한 b1을 비롯한 거대은행들이 가장 형질이 우수하고 순차적으로 매우 형질이 떨어지는 T를 따르는 저축은행 sb1과 sb2가 가장 형질이 떨어진다고 하겠다.

요약하자면, 거대은행들로 이루어진 첫 번째 집단이 가장 우수한 형질을 갖고 있으며 중소은행, 저축은행 순으로 형질이 구분된다.

3.3. 서열 정렬

16개 은행들에 대한 쌍별서열정렬을 수행하였다. 예를 들어, 저축은행 sb1과 sb2의 경우 43번째 문자부터 55번째 문자까지인 'GTATGGTTCATTT'와 'GTATGGTTAATTT'가 가장 유사한 서열로 나타났다. 이는 수익성에 해당하는 '예금경비율, 영업외 손익률, 수지비율, 인건비대총비용비율, 조세공과대총비용비율, 사내유보율, 사내유보대자기자본비율, 적립금비율, 대출채권평균이자율, 예수금평균이자율, 평균배당률, 자기자본배당률, 배당률'로서 이들 지표들 중에서 '수지비율', '대출채권평균이자율' 그



(a) Banks (b) Banks with imaginary banks

Figure 3.2. Tree without roots

리고 '예수금 평균이자율'은 높은 수준이고 그 외는 낮은 수준에 해당한다.

이는 우리가 알고 있듯이 저축은행들은 주로 고객의 예금에 의존하여 운영되고 따라서 높은 예금 이자를 고객에게 제공해 왔고 많은 예금 수입을 올렸다. 그러나 채권은 높은 이자를 보장해야 발행될 수 있는데 그렇지 못하여 배당률 같은 수익성 지표들은 낮은 수준을 보이고 있음이 저축은행의 경영상 어려움을 반영하고 있다. 위와 같은 방법으로 모든 은행, 저축은행 간 지표에 대한 서열정렬을 수행하여 유사한 서열들을 Table 3.2에 정리하였다.

표의 내용을 간단히 정리한다면, (sb1, sb2)는 수익성에서 (sb3, sb4, sb5)는 생산성에서 유사한 서열이 발견되었고 낮은 수준(G, T)이 주로 나타났다. 은행들 중에서는 b1이 (b3, b5)와 수익성과 주로 안정성에서 유사한 서열이 발견되었으나 다른 은행들과는 유사한 서열이 발견되지 않아 b1이 다른 은행들과는 매우 다른 구조를 갖고 있다고 하겠다. (b6, b7, b8, b9, b10)는 수익성과 생산성 두 가지 측면에서 유사한 서열을 갖고 있으면 b11은 홀로 (b7, b10)과 성장성, 생산성 그리고 안정성에서 유사한 서열을 갖고 있었다. 정리하면 (sb1, sb2), (sb3, sb4, sb5), (b1), (b2) 그리고 (b11)이 각각 다소 차별화된 패턴을 보이고 (b3, b4 , b5), (b6, b7, b8, b9, b10)이 각각 유사한 패턴을 보인다고 하겠다.

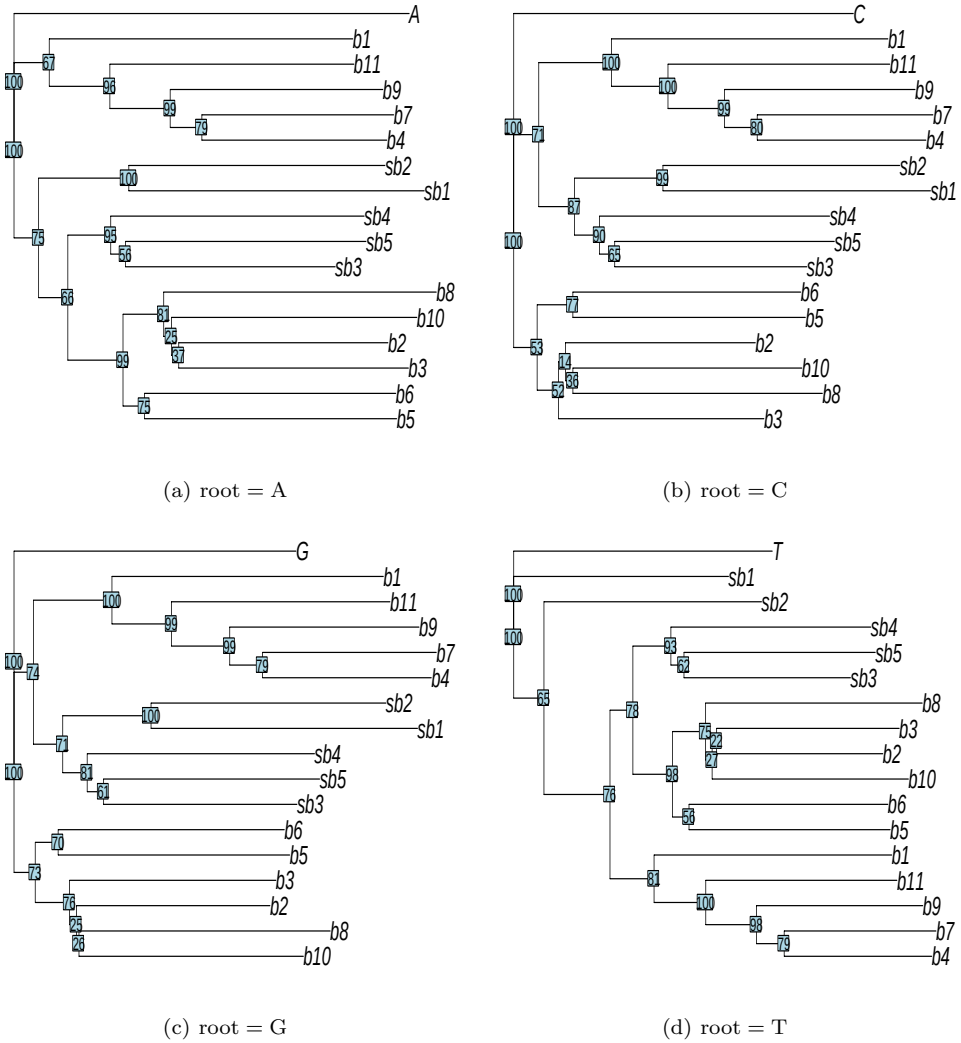
Pak (2013)은 동일한 데이터에 판별분석을 적용하여 은행과 저축은행 간의 판별을 위한 주요 지표로 '조세공과대총비용비율', '대출채권평균이자율'(이상 수익성), '자기자본회전률', '자본금회전률', '예수금회전률'(이상 활동성), '차입금의존도'(안정성)임을 밝혔는데 이는 위의 분석에서 저축은행이 수익성에 관하여, 은행은 활동성과 안정성에 관련된 지표들과 관련이 있다는 사실과 유사하다.

4. 토의

미국 금융 위기가 시작된 2008년의 은행과 저축은행 자료에 대하여 동일한 분석을 수행하였다. 저축은행을 중심으로 대표적인 몇 가지 결과를 보면, 그 결과가 2011년과 매우 유사함을 볼 수 있다. 그런 가

Table 3.2. Common factors for banks; major factors are in parathesis

sb1, sb2	sb3, sb4, sb5	b1, b3, b5
Operating Expenses to Deposits	Value Added	Cashflow Per Share
Non-Operating Income to Operating Revenues	Value Added Per Employee	Book value Per Share
Total Expenses to Total Revenue	Operating Revenues Per Employee	Reserves Ratio
Personnel Expenses to Total Expenses	Operating Income Per Employee	Operating Income Par Value
Taxes & Dues to Total Expenses	Ordinary Income Per Employee (Prior to 2007)	Tangable Assets to Total Assets
Accumulated Earning Ratio	Net Income Per Employee	Loans to Deposits
Accumulated Earning to Stockholder's Equity	Personnel Expenses Per Employee	Stockholder's Equity to Total Assets
Additional Paid-In Capital &	Deposits Per Employee	Debt to Total Assets
Retained Earning to Stockholder's Equity		
The Average Rate of Interest on Loans Receivable	Loans Per Employee	
Decrease in The Average Amount of	Liabilities & Stockholder's Equity Per Employee	
Interest Accrued on Deposits Received		
Dividends to Capital Stock	Ratio of Value Added to Liabilities & Stockholder's Equity	
Dividends to Stockholder's Equity		
Dividends to Net Income		
(profitability)	(productivity)	(profitability)/(stability)
b2, b3, b4, b5	b6, b7, b8, b9, b10	b6, b7, b8, b9, b10, b11
Operating Revenues to Operating Expenses	Value Added Per Employee Growth Rate	Value Added Per Employee Growth Rate
Operating Income to Operating Revenues	Employees Growth Rate	Employees Growth Rate
Net Income to Operating Revenues	Operating Revenues Growth Rate Per Employee	Operating Revenues Growth Rate Per Employee
Total Income to Total Capital	Personnel Expenses Growth Rate Per Employee	Personnel Expenses Growth Rate Per Employee
Operating Income to Total Capital	Operating Revenues to Operating Expenses	Operating Revenues to Operating Expenses
Net Income to Total Capital	Operating Income to Operating Revenues	Operating Income to Operating Revenues
Operating Income to Stockholder's Equity	Net Income to Operating Revenues	
Net Income to Stockholder's Equity	Total Income to Total Capital	
Operating Income to Capital Stock	Operating Income to Total Capital	
Net Income to Capital Stock	Net Income to Total Capital	
Operating Expenses to Operating Revenues	Operating Income to Stockholder's Equity	
Operating Expenses to Deposits	Net Income to Stockholder's Equity	
	Operating Income to Capital Stock	
	Net Income to Capital Stock	
	Operating Expenses to Operating Revenues	
	Operating Expenses to Deposits	
(profitability)	growth/(profitability)	(growth)/profitability
b7, b11	b2, b3, b4	b10, b11
Reserves Ratio	Stockholder's Equity to Total Assets	Value Added Per Employee
Operating Income Par Value	Debt to Total Assets	Operating Revenues Per Employee
Tangable Assets to Total Assets	Total Borrowings & Bonds Payable to Total Liabilities	Operating Income Per Employee
Loans to Deposits	Total Borrowings & Bonds Payable to Total Assets	Net Income Per Employee
Stockholder's Equity to Total Assets	Deposits to Stockholder's Equity	Personnel Expenses Per Employee
Debt to Total Assets	Liabilities Ratio	Deposits Per Employee
Total Borrowings & Bonds Payable to Total Liabilities	Reserves to Total Assets	Loans Per Employee
Total Borrowings & Bonds Payable to Total Assets	Cash Flows from Investing Activities Rates	Liabilities & Stockholder's Equity Per Employee
Deposits to Stockholder's Equity	Total Assets Turnover	Ratio of Value Added to Liabilities &
		Stockholder's Equity
Liabilities Ratio	Stockholder's Equity Turnover	Ratio of Value Added to Net Sales
Reserves to Total Assets	Capital Stock Turnover	Ratio of Personnel Expenses to Value Added
	Debt Turnover	1 Minus Personnel Expenses to Value Added
	Loans Turnover	
	Deposits Turnover	
profitability/(stability)	(stability)/(activity)	(growth)

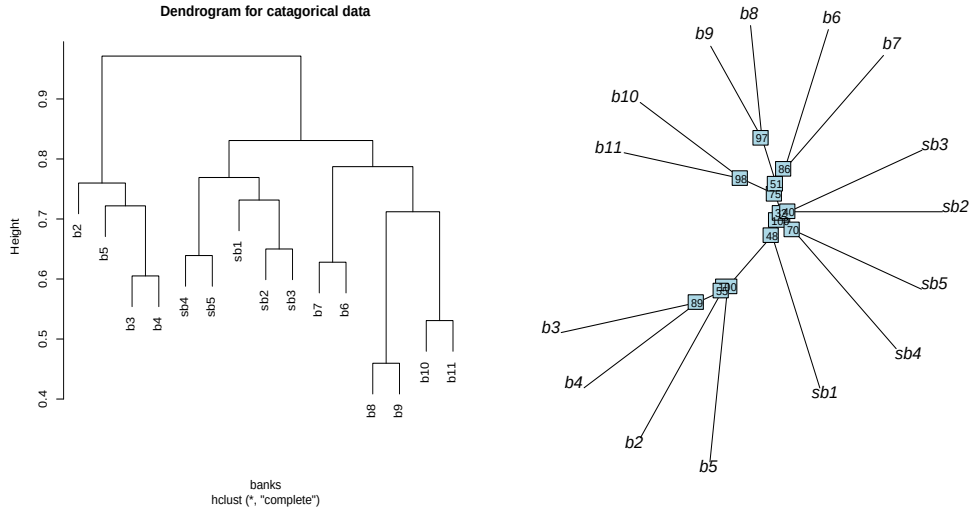


(a) root = A (b) root = C

(c) root = G (d) root = T

Figure 3.3. Tree with roots

운데 눈에 띄는 것은 저축은행 sb1이다. 저축은행 sb1이 2008년에는 다른 저축은행들과 분리된 모양을 볼 수 있다 (Figure 4.1). 2008년도의 경우, 덴드로그램을 보면 (sb1), (sb2, sb3) 그리고 (sb5, sb6)로 묶이고 진화나무에 의하면 sb1은 거대은행에 해당하는 (b2, b3, b4, b5)의 줄기에 연결되어 있음을 볼 수 있다. 우연의 일치인지 모르겠지만 퇴출의 순서, sb3 (2011년 9월) → (sb2, sb5) (2012년 5월) → sb4 (2012년 11월) → sb1 (2013년 2월)와 진화나무의 원점에 가까운 순서가 매우 유사함을 발견할 수 있다. 그런대로 은행들과 유사한 패턴을 보인 sb1이 가장 늦게 퇴출 되었다. 즉, 진화의 최초 원점에 가까운 저축은행부터 퇴출 된 것은 발전이 덜 된 혹은 변화가 없었던 저축은행부터 퇴출 되었다는 추리를 가능하게 한다. 2008년도에 b1은 지주회사 전환 과정에서 결측 데이터가 많이 발생하여 분석에서 빠져



(a) Clustering (b) Unrooted tree

Figure 4.1. Results with data for year 2008.

있다.

5. 결론

최초로 수집된 자료의 형태가 연속형이었으나 분석이 여의치 않아 연속형 자료를 범주형 자료로 바꾸어 분석을 시도하였다. 이 경우 정보 손실의 우려가 있을 수 있으나 어쩔 수 없는 경우라면 범주형으로 변환한 경우 얻을 수 있는 효과가 존재함을 보였다. 본 논문에서 제시한 방법이 기존의 방법들보다 우수함을 보이기에는 군집분석의 특성상 한계가 있으나, 연속형 분석이 가능하지 않을 때 대체적인 방법으로 시도할 수도 있음을 보이려 하였다. 본 연구에서는 국내 은행과 저축은행 자료를 예제로 분석함에 있어 범주화를 통하면 범주형 자료 분석 특유의 기법을 통해 탐색적 목적으로 혹은 연속형 자료 분석이 제시못하는 나름의 유용한 지식을 획득할 수 있음을 보였다. 자료를 분석한 결과 거대은행, 중소은행 그리고 저축은행이 특정한 지표에서 높은 혹은 낮은 값을 구별되게 갖고 있음을 확인할 수 있었다.

References

Baldauf, S. L. (2003). Phylogeny for the Faint of Heart: A tutorial, *Trends in Genetics*, **19**, 345–351.

Barry, D. and Hartigan, J. A. (1987). Asynchronous distance between homologous DNA sequences, *Biometrics*, **43**, 261–276.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution*, **17**, 368–376.

Fitch, W. M. (1966). Mutation values for the interconversion of amino acid pair, *Journal of Molecular Biology*, **16**, 9–16.

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

Jin, L. and Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis, *Molecular Biology and Evolution*, **7**, 82–102.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of Protein Molecules. In Mammalian Protein Metabolism*, ed. Munro, H. N., Academic Press, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *Journal of Molecular Evolution*, **16**, 111–120.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences, *Proceedings of the National Academy of Sciences USA*, **78**, 454–458.

Krane, D. E. and Raymer, M. L. (2003). *Fundamental Concepts of Bioinformatics*, Pearson Education, San Francisco, CA.

Olsen, G. J. (2013). Phylogenetic Analysis, *Course Handout*, Available from: http://www.life.illinois.edu/mcb/432/Handouts/PhylogeneticAnalysis.pdf.

Pak, R. J. (2013). Key financial indexes classifying banks and savings banks, *Journal of the Korean Data Analysis Society*, **15**, 719–730.

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

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, *Molecular Biology and Evolution*, **10**, 512–526.

은행과 저축은행 관련 재정 지표 분석: 생물 정보학 분석 기법의 응용

박노진[a,1]

[a]단국대학교 응용통계학과

(2014년 4월 17일 접수, 2014년 5월 21일 수정, 2014년 6월 18일 채택)

요 약

자료의 수집과 저장이 수월해 지면서 대용량의 자료들이 존재하고 특히 개체 보다 변수가 더 많은 자료들이 생산되고 있다. 변수들이 증가하면서 다중공선성 같은 문제들이 발생하여 분석의 어려움에 봉착하게 된다. 이러한 문제를 해결하는 방법들이 많이 연구되었지만 다소간의 정보의 손실을 감내하고 연속형 자료를 범주형 자료로 변환하면 나름 유용한 분석이 가능하다고 본다. 대용량 범주형 자료의 대표적인 사례로 유전자 염기 서열 자료가 있고 이를 분석하기 위한 많은 기술들이 발달되어 있다. 본 논문에서는 국내 은행들이 생산해 낸 다양한 지표들을 분석하기 위해 유전자 염기 서열 분석 기법을 적용하여 분석하였고 나름 유용한 정보를 얻을 수 있음을 보였다. 본 논문에서 사용한 자료는 11개의 은행과 5개의 저축은행과 관련된 78개 재정 지표를 갖는 자료로서 심각한 다중 공선성이 존재하여 자료를 범주화하고 분석한 결과 몇 가지 유용한 결과를 도출하였다.

주요용0어: 다중 공선성, 범주형 자료 분석, 유전자 서열 분석.

본 연구는 2014학년도 단국대학교 대학연구비 지원으로 연구되었음.

[1](448-701) 용인시 수지구 죽전동 126, 단국대학교 응용통계학과. E-mail: rjpak@dankook.ac.kr