시스템 R을 활용한 범죄데이터 분석 기법 연구

장우인[1], 오재섭[2], 박영호[1,*]
[1]숙명여자대학원 멀티미디어과학과,
[2]숙명여대 정책산업대학원
[*]교신저자
e-mail: leader0710@sm.ac.kr, well72@gmail.com, yhpark@sm.ac.kr

A Study on Crime Data Analysis Technique Using System R

Wu-In Jang[1], Jae-Suhp Oh[2], Young-Ho Park[1,*]
[1]Dept of Multimedia Science, Sookmyung Women's University
[2]Graduate School of Public Policy and Industry,
Sookmyung Women's University
[*]Corresponding Author

요 약

최근 SNS나 스마트폰을 이용한 다양하고 많은 데이터들이 우리 주위에 산재하고 있다. 이러한 데이터는 사용자의 심리나 상황을 담은 것으로 이에 대한 분석을 통해 사용자의 행동과 심리를 유추할 수 있다. 그러나 이러한 데이터는 빅데이터의 형태를 가지고 있기 때문에 이를 효과적으로 분석하기 위한 방법론이 필요하다. 본 논문에서는 이러한 문제에 초점을 맞추고 이를 효과적으로 분석하기 위하여, 먼저 시스템인 R을 소개하고, R에 실제 데이터를 로딩 하여, 이를 분석하는 분석 예를 보인다.

1. 서론

최근 몇 년간 스마트폰의 빠른 보급과 블로그 및 페이스북과 같은 소셜 미디어의 활성화는 데이터의 양적, 질적인 팽창을 가져왔다. 시장 조사 업체인 IDC가 발표한 "디지털 유니버스 보고서(IDC= Digital Universe Study)"[1]에 따르면 2011년 전 세계에서 생성되는 디지털 정보량이 1.8 제타바이트에 달하고, 전 세계의 디지털 정보량은 매 2년마다 2배씩 증가한다고 한다. 여기서 1.8 제타바이트는 대한민국의 모든 사람(4858만 명, 2010년 기준)들이 17만 847년 동안 쉬지 않고 매 분마다 3개의 트위터 글을 게시한 양[1]과 같다. 이렇듯 최근 몇 년 사이에 기업에서 처리해야 하는 데이터의 양은 급격히 증가했으며 이러한 빅데이터를 효과적으로 분석해야 하는 방법론이 필요하게 되었다.

빅데이터 처리의 문제는 많은 연구에서 검색 및 저장 성능, 검색 편의 및 유용성, 의미 검색 등에 초점을 맞추어 다양한 방법으로 해결해 왔다. 본 논문에서는 데이터 처리의 유용성 측면에서 각광을 받고 있는 소프트웨어 프레임워크인 Map-Reduce[2]와 Hadoop[3]을 활용하여 해결 방안을 찾은 기존 연구들에 관하여 간략히 소개하고 범죄 데이터를 가지고 시스템 R[4]을 활용하는 방안을 고찰해 보고자 한다.

2. 관련연구

[5]에서는, 빅데이터를 효율적으로 분석 및 마이닝을 수행하기 위하여 Map-Reduce 기술에 대한 연구가 활발히 진행되고 있음을 밝히고 특히 순환 처리 응용을 위하여 Map-Reduce 기술이 필요함을 주장한다. 다양한 순환 처리 응용을 통한 성능평가를 수행하여 각 분산 시스템에 대한 성능을 비교함으로써 각 순환 처리 응용에 적합한 Map-Reduce 기반 분산 시스템의 요구사항을 분석한다.

[6]에서는, 빅데이터를 바라보는 시각을 크게 기술적 관점과 분석적 관점으로 나누고 특히 기술적 관점에서 Hadoop을 표준으로 하는 오픈소스 분석 플랫폼에 초점을 맞춘다. Hadoop 플랫폼의 구성을 설명함으로써 그 자체만으로는 완벽하지 않음을 밝히고 이를 위해 NoSQL과 RDBMS, 메모리 기반의 캐시 시스템을 함께 사용하는 활용방안을 제시한다.

3. 시스템 R

R은 통계학자들이 만든, 통계학자들을 위한 통계 프로그램이다. 통계 계산과 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경이다[7]. 이 언어는 AT&T에서 개발한 분석용 언어인 S에서 영향을 받아 개설되어 GNU S라는 이름으로 불리기도 한다[8]. R은 사용자들에게 무료로 제공되는 소프트웨어다. 누구나 자유롭게 실행, 복사, 수정, 배포 할 수 있고, 누구도 그런 권리를 제한하면 안 된다는 사용 허가권 아래 소프트웨어를 개발, 배포[9]하였다. 이것

은 자유를 지향하는 R의 철학을 담고 있다. 다른 개발언어와는 달리 대부분 인식성이 높은 함수 이름을 사용하여 비전공자들도 기본 교육을 받으면 사용이 가능하다.

3.1 R의 특징

R을 표현해 줄 수 있는 키워드는 데이터, 통계, 그래픽이다. 데이터는 테이블 형태, 혹은 열과 행으로 구성된 스프레드시트 형태로 구성되어 있다. *read()* 함수를 이용하거나 엑셀파일을 이용해서 로딩하고 저장하며 묶기, 정렬과 같이 데이터들을 활용할 수 있다. R과 SAS, SPSS와 같은 기존의 통계분석 소프트웨어의 가장 큰 차이점은 R은 오브젝트 기반 객체 지향적 언어라는 것이다. SAS의 '프로시저'가 아닌 함수 중심으로 분석가가 분석 로직을 'R스크립트'를 이용해 자유자재로 구현할 수 있다. List, Array, Vector 등과 같이 연산과 관련된 다양한 데이터 구조를 제공하고, 그것으로부터 다양한 통계 이론을 적용시킬 수 있다. 그래픽은 R의 장점이다. 그래픽스 패키지를 통해 분석에 적합한 그래픽스를 제공해준다. 디자인에서부터 그래프의 종류까지 다양한 그래픽을 만드는 데 필요한 유용한 함수들로 구성되어 있는 것이 그래픽스 패키지이다. 그래픽을 통하여 예측, 분석, 가설검증 등이 가능하게 된다. 특히 빅데이터와 관련한 R의 장점은 인메모리(in-memory)방식이라 Map-Reduce 방식을 적용하기 쉽고, Hadoop과 연동되는 패키지들[8]이 존재한다는 것이다. 그러므로 기존 통계 프로그램 패키지들과는 차별화된 분석 기능을 제공해준다.

이러한 편리성과 기능들을 바탕으로 Google, Facebook, ACR, ANZ, bing 등의 IT 기업들이 R을 분석 플랫폼, 분석 엔진[10]으로 활용하고 있다.

3.2 일반적인 기능처리 방법

본 내용에 들어가기에 앞서 [11]을 참조하여 다음을 정리하였음을 밝힌다. R을 실행하기에 앞서 기본적인 사항을 기술하겠다. 가장 먼저 R을 컴퓨터에 설치한다. 윈도와 OS X 사용자들은 Comprehensive R Archive Network의 약자인 CRAN에서 R을 다운로드 할 수 있다. 리눅스와 유닉스의 사용자는 패키지 관리 도구를 통해 R 패키지를 설치할 수 있다. 이렇게 설치된 R을 실행한다. R을 실행하면 새 창이 열린다. R 콘솔 창에 R식을 입력할 수 있다. R에서 '>'는 프롬프트라고 한다. 프롬프트 다음에 표현식을 입력하면 R이 값을 평가해 결과를 화면에 보여주게 된다. 작업을 하는 중간에라도 R의 작업공간을 저장하고 싶다면 *save.image()* 함수를 호출해서 사용하면 된다.

R의 주요 기능들 중 대표적으로 데이터 로딩, 확률과 통계, 그래픽스를 처리하는 방법을 살펴보고자 한다. 첫 번째는 다루고자 하는 데이터를 로딩 하는 과정이다. 데이터를 로딩하기에 앞서 먼저 각각의 데이터를 셀 안에 삽입하는 엑셀작업을 수행하면 좋다. 엑셀을 사용하는 이유는 데이터들을 구분하기에 용이하고, 시스템 R에서 엑셀파일을 지원하기 때문이다. 이 엑셀파일을 시스템 R에 로딩을 하고자 하면 *read.table()* 함수를 이용하면 된다. txt 파일 형식으로 저장을 했다면 *read.table("파일이름.txt", header=T)*로 입력하면 외부 데이터로부터 데이터 프레임을 생성, 로딩 할 수 있게 된다.

두 번째는 확률과 통계를 구하는 방법이다. 이것들을 다루는 함수들이 많아서 몇 개만 소개하도록 하겠다. 우선 확률에서는 이산변수들의 확률을 계산할 때 자주 접할 수 있는 조합의 개수 세기를 계산해 주는 함수로는 *choose()*가 있다. 여기에 조합을 만들어 내기 위해서는 *combn()* 함수를 사용한다. 재현 가능한 난수를 생성할 땐 *set.seed()* 함수를 호출하면 된다. 통계에서는 데이터에 관한 기본적인 통계 정보들을 요약하고 싶을 때 *summary()* 함수를 사용한다. *summary()* 함수는 벡터, 행렬, 요인, 데이터 프레임에 대한 통계량을 보여준다. 데이터세트가 하나 있고 모든 데이터 원소에 대해 각각에 상응하는 z 점수를 계산하고 싶을 땐 *scale()* 함수를 호출하면 된다.

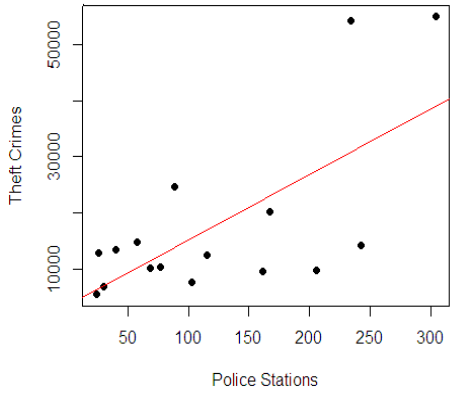
세 번째는 로딩한 데이터를 활용하여 두 변수간의 상관관계를 밝히는 그래픽스를 나타내는 과정이다. 우선 비교하고자 하는 데이터를 x축과 y축에 설정을 한다. *cor()* 함수를 사용하여 둘의 상관계수를 구한 뒤, *plot()* 함수의 pch 인자를 사용하여 산점도를 나타낼 수 있다. 이 때 *x*축과 *y*축에 lab 인자를 사용하면 라벨을 추가할 수 있고, pch 인자에 19를 대입하면 흰색이었던 점들이 검은색으로 바뀌게 된다. 예를 들자면 *plot(x, y, xlab="Police Stations", ylab="Violent Crimes", pch=19)* 이런 식으로 입력하면 된다. 점들의 선형회귀를 보여주고 싶다면 회귀선을 그려주는 *abline()* 함수와 *lm()* 함수를 사용하면 된다. 선의 색상도 지정할 수 있는데 col="색상"으로 선언하면 원하는 색상으로 변경이 가능하다. 이렇게 완성된 그래픽을 지우고 싶다면 *dev.off()* 함수를 사용하면 된다.

4. 범죄 데이터

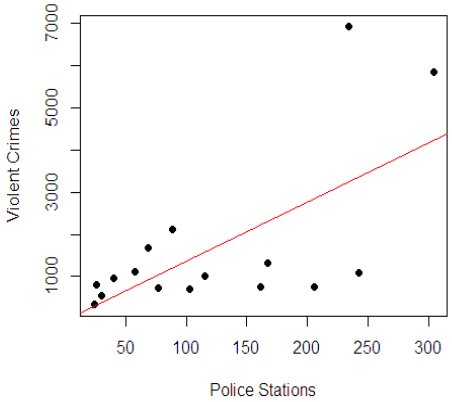
간단한 실험을 통해서 R을 활용해 보고자 한다. 숫자로 표현할 수 있는 데이터를 다루기에 편리한 R의 특성상 경찰청에서 공개한 실제 통계 자료를 가지고 실험을 했다. 2011년에 지역별로 발생한 범죄 데이터와 16개 시도별 경찰 인력에 대한 데이터를 활용하여 이 둘의 상관관계를 밝히고자 한다. 범죄유형은 다른 범죄들도 있지만 대한민국에서 가장 많이 발생한 강력 범죄(살인, 강도, 강간, 강제추행, 방화)와 절도 범죄로 국한하였고, 경찰 인력과는 별도로 경찰서의 숫자를 포함하여 각각 데이터로 삼았다.

5. 시스템 R을 활용한 범죄 데이터 분석

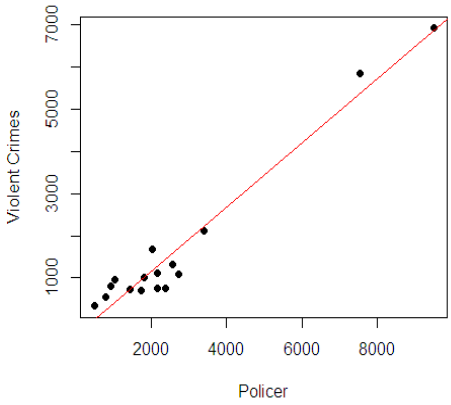
경찰인력이 많을수록 범죄 예방 효과가 있다는 가설이 있다. 범죄 유형과 경찰 인력간의 상호 관련성을 통하여 이 가설이 사실인지 거짓인지를 증명해 보이도록 하겠다. 범죄 유형과 경찰 인력과의 상관관계는 그림 1~4를 통해서 알 수 있다. 그림 1의 x, y축은 경찰서 수와 절도 범죄



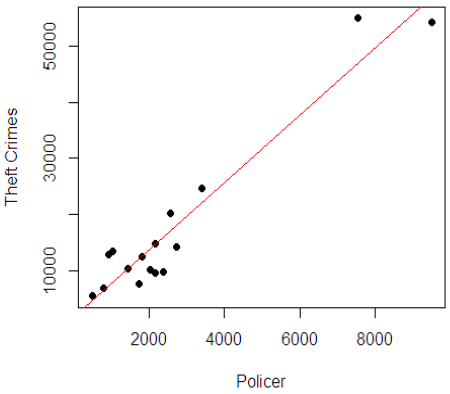
(그림 1) 경찰서와 절도 범죄간의 상관관계 (r= 0.675)



(그림 2) 경찰서와 강력 범죄간의 상관관계 (r= 0.646)



(그림 3) 경찰관과 강력 범죄간의 상관관계 (r= 0.977)



(그림 4) 경찰관과 절도 범죄간의 상관관계 (r= 0.960)

발생건수, 그림 2의 x, y축은 경찰서 수와 강력 범죄 발생건수, 그림 3의 x, y축은 경찰관 수와 강력 범죄 발생건수, 그림 4의 x, y축은 경찰관 수와 절도 범죄 발생건수로 각각 설정했다. *cor(x, y)* 함수를 이용하여 두 변수간의 상관계수를 나타냈다. 그림 1은 0.6747387, 그림 2는 0.6456997, 그림 3은 0.97685, 그림 4는 0.9598508로 그림 3이 가장 높은 상관계수를 갖는다.

검은 점은 산점도를 나타낸다. 산점도는 x와 y가 서로 관계가 있다는 가정 하에, 그 관계를 파악하기에 좋다[12]. 산점도의 목적은 *plot()* 함수를 이용하여 x와 y, 둘의 관계를 그래픽스 창에 점들로 나타내는 것이다. 상관계수 절대값이 클수록 직선에 점들이 모이게 된다.

abline() 함수를 사용해서 회귀선을 추가했다. x와 y간의 상관관계가 높을수록 가파른 회귀선이 그려진다. 그림 1과 2를 묶고, 그림 3과 4를 묶어서 비교해보면 경찰서의 숫자보다는 경찰관의 숫자와 강력 범죄, 절도 범죄에 해당하는 데이터가 밀접한 관계를 가지고 있는 것으로 나타났다. 이를 통해 5장 처음에 명시한 가설은 거짓으로 범죄가 많이 발생하는 지역에 경찰 인력이 많다는 점을 증명 했다.

6. 결론

본 논문에서는 빅데이터 처리에 대한 문제에 초점을 맞추고 이를 효과적으로 분석하기 위하여 시스템 R을 이용했다. 시스템 R을 통해서 데이터들 간의 상관관계를 밝히는 분석 결과를 보였다. 범죄발생건수와 경찰인력의 숫자는 정비례하다는 결과를 얻었다. 이를 통하여 범죄 예방 효과를 기대하고자 경찰인력의 숫자를 늘리는 해결방안은 효과적이지 못함을 증명 했다. 경찰인력 숫자에 초점을 맞추는 단순한 접근 방법이 아니라, 보다 근본적인 다른 방안이 필요하다는 것을 나타내기 때문에 이 결과는 의미가 있다.

시스템 R을 사용하여 간단한 제네릭 그래프 함수를 만들고 x, y 두 변수 간에 관계를 밝혔다. 이것은 사회현상을 분석하고 원인과 결과를 밝혀주는 상호 관계성을 파악하는데 유용하다. 때론 우리가 흔히 알고 있는 통념이 잘못되었음을 증명하는 역할을 하기도 한다. 향후에는 시스템 R에 포함된 기능들을 여러 가지 시도를 통해 지금보다 더 확장된 실험 결과를 로딩해 보겠다. 더 나아가 R의 단점을 보완해 줄 수 있는 새로운 시스템을 제안할 것을 계획한다.

7. 사사문구

본 연구는 지식경제부 및 한국산업기술평가관리원의

산업융합원천기술개발사업의 일환으로 수행하였다. 과제번호와 과제명은 다음과 같다. 10041854, 안전한 주거환경을 위한 실시간 위험요소 예측/방지용 스마트 홈 서비스 플랫폼 기술 개발.

참고문헌

[1] 김재훈, “빅데이터 시대 도래에 따른 데이터 처리기술 현황과 전망,” 이슈리포트 제 21호, 2011.
[2] Map-Reduce, http://hadoop.apache.org/docs/stable/mapred_tutorial.htm
[3] Hodoop, http://hadoop.apache.org/
[4] R 홈페이지, http://www.r-project.org/
[5] 장성재, 김형일, 윤민, 최동훈, 조희승, 장재우, “빅데이터 순환 처리 응용을 위한 Map-Reduce 기반 분산 시스템의 성능평가,” 2013 한국컴퓨터종합학술대회 논문집 pp.334~336, 2013.
[6] 이현종, “빅데이터 하둡 플랫폼의 활용,” 한국통신학회 논문지 제 29권 제11호, pp.43~47, 2012.
[7] 위키백과, “R 프로그래밍 언어,” http://ko.wikipedia.org/wiki/R_(%ED%94%84%EB%A1%9C%EA%B7%B8%EB%9E%98%EB%B0%8D_%EC%96%B8%EC%96%B4)
[8] 아크원소프트 공식블로그, “빅데이터와 R,” http://blog.naver.com/qrrmaa112?Redirect=Log&logNo=120191365636
[9] Jeon Hee-Won, “오픈소스 기반의 통계언어 R과 빅데이터 분석,” KTNexR R 실무 데이터 분석 세미나.
[10]REVOLUTION ANALYTICS, “Companies Using R,” http://www.revolutionanalytics.com/what-is-open-source-r/companies-using-r.php
[11] 폴티터, R cookbook. 인사이트, 1장, 8장~10장, 2012.
[12] 폴티터, R cookbook. 인사이트, pp.281, 2012.