Journal of The Korea Society of Computer and Information
Vol. 17 No. 4 , April 2012

2012-17-4-5-3

범죄유형별 범죄발생 예측확률을 높일 수 있는 방법에 관한 연구

정 영 석*, 김 진 묵**, 박 구 락*

A study of improved ways of the predicted probability to criminal types

Young-Suk Chung*, Jin-Mook Kim**, Koo-Rack Park*

요 약

현대 사회는 다양한 강력 범죄들이 발생하고 있다. 모든 범죄들은 발생한 후에 대처를 하는 것보다 사전에 범죄를 예방하는 것이 가장 중요하다. 이를 위해서 다양한 범죄를 예방하기 위한 연구가 진행되었다. 하지만 기존 연구 방법들은 사회학적, 심리학적인 요인들을 분석하여 범죄의 발생 확률과 발생 동기 등을 분석하여 예방하고자 하는 노력이 대부분이다. 그러므로 본 논문에서는 마코프 체인 방식을 사용하여 시간에 따른 범죄를 예측하기 위한 연구를 수행하고자 한다. 5대 강력 범죄인 강도, 살인, 강간, 절도, 폭력에 대하여 수집된 범죄 발생 건수 자료를 사용해 범죄 유형별 시간에 따른 범죄 발생 예측을 위한 모델링을 구현한다. 그리고 범죄 발생 유형별 범죄 발생 예측 값과 실제 발생 값을 비교해 본 논문에서 제안한 시간에 따른 범죄 발생 예측 모델링의 타당성을 검토하였다. 본 논문에서 제안한 범죄 발생 예측 기법이 실제로 강도, 살인, 강간 등과 같은 강력 범죄에 대해서는 최대 값을 임계값으로 적용하고, 나머지 범죄에 대해서는 평균값을 적용하는 방식을 사용함으로써 범죄 발생 예측확률을 높일 수 있을 것으로 연구되었다. 향후 범죄 유형별로 시간에 따른 범죄발생 예측율과 실제 값이 다르게 적용되는 사례들을 추가 조사하여 연구의 폭을 넓히고자 한다.

▸ Keyword : 시뮬레이션, 범죄 통계, 미래 예측, 마코프 체인

Abstract

Modern society, various great strength crimes are producing. After all crimes happen, it is most important that prevent crime beforehand than that cope. So, many research studied to prevent

• 제1저자 : 정영석 • 교신저자 : 박구락
• 투고일 : 2011. 12. 16, 심사일 : 2012. 01. 29, 게재확정일 : 2012. 02. 26.
* 공주대학교 컴퓨터공학과(Dept. of Computer Science & Engineering, Kongju national University)
**선문대학교 IT 교육학부(Dept. of IT Education, Sun moon University)

various crime. However, existing method of studies are to analyze and prevent by society and psychological factors. Therefore we wishes to achieve research to forecast crime by time using Markov chain method. We embody modelling for crime occurrence estimate by crime type time using crime occurrence number of item data that is collected about 5 great strength offender strength, murder, rape, moderation, violence. And examined propriety of crime occurrence estimate modelling by time that propose in treatise that compare crime occurrence type crime occurrence estimate price and actuality occurrence value. Our proposed crime occurrence estimate techniques studied to apply maximum value by critcal value about great strength crime such as strength, murder, rape etc. actually, and heighten crime occurrence estimate probability by using way to apply mean value about remainder crime in this paper. So, we wish to more study about wide crime case and as the crime occurrence estimate rate and actuality value by time are different in crime type hereafter applied examples investigating.

▸ Keyword : Simulation, Crime statistics, Predictive modeling, Markov chains

I. 서 론

산업과 과학기술의 발전으로 인류의 삶의 질이 향상되고 있는 반면에 실업, 빈곤, 범죄 등 다양한 문제들이 발생 하고 있다. 그 중 범죄는 인간이 활동하고 거주하는 모든 지역에서 발생하고 있다고 해도 과언이 아니다. 범죄로 인한 피해도 증가 추세에 있다. 그러므로 범죄에 대한 다양한 연구가 이루어지고 있다. 예를 들면, 범죄자와 피해자의 거주지 사이에 공간적 분포의 패턴을 연구하여, 특정한 패턴의 원인을 찾는 것이 연구 되고 있고[1][2], 도시공간구조의 특성을 파악해서 범죄 발생 가능성을 평가 하는 연구가 있었고[3], 범행 위치의 공간적 분포와 범죄 발생 의 시간적 분포 특성에 따라 시공간 패턴을 유형화 하여 연쇄 범죄의 위치를 예측하는 연구가 있었다[4]. 그리고 범죄 경력을 바탕으로 범죄자들이 유사한 범죄를 지속적으로 저지르는지, 아니면 다양한 범죄들을 저지르는 지에 대한 범죄 유형전환에 관한 연구도 있다[5]. 그러나 기존의 연구방법들은 지리적, 심리적, 사회학정인 요인을 분석하는 연구가 대부분이었고, 시간에 따른 범죄 발생 건수를 예측하는 연구는 부족하다. 그래서 본 논문에서는 마코프 체인 이론을 이용하여 범죄 발생 건수를 예측 할 수 있는 연구를 진행하고자 한다. 그리고 범죄 발생 건수 예측을 높이는 방법에 대해 논의한다. 본 논문의 구성은 다음과 같다. 2장에서 관련연구인 마코프 체인에 대해 알아보고 3장에서는 마코프 체인을 이용한 범죄 예측 모델링에 대해 제안한다. 4장에서는 범죄 예측 모델링을 바탕으로 실제 범죄 데이터를 이용하여 예측하고 그 값을 실제 발생 건수와 비교해 본다. 5장에서는 결론 및 향후 연구 과제에 대해 논의 한다.

II. 관련 연구

1. 마코프 체인

마코프 체인은 마코프 성질을 가진 이산 확률과정으로서 러시아의 수학자인 안드레이 마코프의 이름에서 왔다. 마코프 체인은 어떤 사건에서 직전의 상태가 현재의 상태로 영향을 주고, 과거의 상태에는 전혀 영향을 받지 않는 확률 과정을 가정한 것이다. 마코프 체인은 과거의 동적 특성을 분석하여 미래에 있을 변화를 예측하기 위한 수학적 기법이다.[6] 마코프 체인을 날씨를 예로 들어 설명하면 오늘의 날씨가 맑았다고 해서 내일의 날씨도 맑다고 할 수 없다. 확률적으로는 맑은 날, 흐린 날, 비 가오는 날이 될 수도 있다. 이것을 비결정 시스템이라고 하고 이러한 문제를 해결하기 위해 사용된다[7].

임의의 시간 $t_1 < t_2 < \cdots < t_k < t_{k+1}$ 에 대해 $X(t)$가 이산 값을 나타낼 때를 마코프 체인이라하고 하면

$$P[a < X(t_{k+1}) = x_{k+1} | X(t_k) = x_k, \cdots, X(t_1) = x_1] \quad (1)$$
$$= P[X < X(t_{k+1}) = x_{k+1} | X(t_k) = x_k]$$

식(1)로 기술할 수 있다. 위의 식(1)에서 t_k는 현재, t_{k+1}

은 미래, $t_1, \cdots, t_{k-1}$은 과거의 시점이다. 마코프 체인은 가능한 상태들의 집합인 상태집합, 초기화 확률 벡터인 초기 확률, 각 상태간의 전이로 구성된 전이 행렬로 구성되어 있다[8]. 미래를 예측할 수 있는 마코프 체인의 특성은 다양한 분야에서 사용되고 있다. 마코프 체인을 이용하여 정류장 사이를 전이 확률을 이용한 행렬 표를 생성하여 버스지체 시간을 예측하기 위한 연구가 있었고[9], 공동주택내의 쾌적한 실내 환경을 유지하기 위해 재실자의 행동을 확률적으로 예측하는 연구에 사용 되고 있다[10]. 또한 전력 계통 시스템의 과거 고장데이터를 이용하여 다음에 발생될 전력기기의 고장 건수를 예측하는 모델로 마코프 체인을 이용[11]하는 등 다양한 분야에 활용되고 있다.

III. 본 론

1.범죄 예측 과정

기존 연구에 있어서 기존의 마코프 체인 이론은 시간에 따른 상태의 변화를 예측할 때 어떤 상태에서의 중간 값 또는 최근 발생한 사건의 평균 값 만을 이용하였다. 그러나 이런 경우 중간 값 또는 평균 값 이상을 예측할 수 없다. 본 논문에서는 중간 값으로 마지막 5 개월간의 평균 값 외에 통계 기간 중 발생한 사건의 최대 값을 적용하여 범죄 발생 건수의 예측 값을 산출하려 한다.

1.1 범죄예측 모델링

마코프 체인을 이용한 범죄 예측 모델링을 보면 그림 1과 같다.

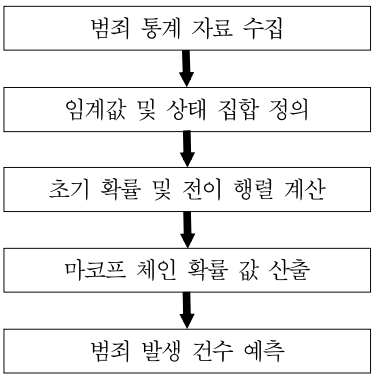

그림 1. 마코프 체인을 이용한 범죄 예측 모델링
Fig. 1Crime prediction modeling using Markov chain

대검찰청이 매년 발표하는 범죄 발생 통계 자료를 1년 단위로 수집하여 임계값을 설정하고, 임계값을 기준으로 상태집합을 정의한다. 정의된 상태 집합으로 초기 확률과 전이행렬을 생성하고, 이것을 마코프 체인을 이용한 범죄 예측 모델링에 입력하여 범죄 발생 예상 건수를 예측 한다.

1.2 범죄 예측 모델링 구현

대검찰청에서 발표하는 범죄 통계 자료 중 5대 범죄(강도, 살인, 강간, 폭력, 절도)를 대상으로 범죄 발생 건수를 예측하였다. 마코프 체인은 임계값을 바탕으로 상태집합을 정의하고, 초기 확률, 전이 행렬을 계산한 후, 초기 확률과 전이 행렬을 이용하여 마코프 체인 확률 값을 구한다[7,8].

◆ 상태집합(S): 범죄 예측 모델링에서 상태 란 한 범죄가 발생하는 빈도수의 범위 말하며, 범죄 통계 자료를 바탕으로 적절한 임계값으로 설정한 상태들을 집합으로 정의 하였다.

◆ 초기 확률: 한 범죄가 초기상태에서 가질 수 있는 범죄 발생 확률로서, 최근에 발생한 범죄의 발생 상태를 이용하여 식(2)과 같이 정의한다.

$$P(S_1, S_2, \cdots, S_n) = P(\frac{a}{F}, \frac{b}{F}, \cdots, \frac{c}{F}) \quad (2)$$

여기서 $a, b, c,$는 각 상태$(S_1, S_2, \cdots, S_n)$의 범죄 발생 횟수이고 F는 $a, b, c,$의 합이다.

◆ 전이행렬: 정의된 상태간의 전이 상태를 확률로 나타낸 것이다. 각 범죄 발생 데이터들을 상태 집합과 매칭 하여 상태들을 열거한다. 그리고 열거된 하나의 상태에서 다른 상태로의 전이 회수를 구한 후, 이를 전이 행렬로 나타낸 것이다. 각 열은 하나의 상태에서 다른 상태로의 확률을 나타내며, 각 행의 합은 1인 조건을 만족한다. 식(3)의 P는 전이행렬로서 조건(4)을 만족한다.

$$P = \begin{pmatrix} P_{11} & P_{12} & P_{13} & \cdots & P_{1n} \\ P_{21} & P_{22} & P_{23} & \cdots & P_{2n} \\ \cdots & \cdots & \cdots & P_{ij} & \cdots \\ P_{n1} & P_{n2} & P_{n3} & \cdots & P_{nn} \end{pmatrix} \quad (3)$$

$$\sum_{j=1}^{n} P_{1j} = 1,\ \sum_{j=1}^{n} P_{2j} = 1,\ \cdots,\ \sum_{j=1}^{n} P_{nj} = 1,\ P_{ij} \geq 0$$

$$\sum_{j=1}^{n} P_{ij} = 1,\ i = 1, 2, \ldots, n \quad (4)$$

◆ 범죄 발생 확률: 마코프 체인으로 식(5)로 정의 된다.

$$P(S_k) = \sum_{i=1}^{n} P(S_i) P_{ik} \quad (5)$$

$P(S_i)$: 초기 확률 P_{ik}:전이행렬

◆ 범죄 발생 예측 건수 : 본 논문에서는 범죄예측 건수를 예측할 때 평균 값, 최대 값을 이용해서 계산하고 비교해보려 한다.

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, M(S_i) \quad (6)$$

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, Max(S_i) \quad (7)$$

n: 범죄발생 상태 집합의 상태 수

$P(S_i)$: 범죄 발생 확률,

$M(S_i)$: 각 범죄 발생 건수의 평균 값

(최근 5개월간 발생한 범죄의 평균 발생 건수)

$Max(S_i)$: 각 범죄 발생 건수의 최대 값

1.3 범죄 발생 모델링 적용

현재 까지 이루어진 마코프 체인을 이용한 미래 예측 연구에는 각 위협상태의 평균 값[7], 또는 중간 값[8]을 이용하여 미래를 예측하였다. 그러나 이 경우에는 평균 값, 또는 중간 값 이상의 값을 예측 할 수 없다. 그래서 본 논문은 평균 값, 최대 값을 적용하여 어느 값이 더 범죄 예측에 가까운지 비교해 보려고 한다.

1.3.1 범죄 유형 (강도)

표 1은 2006~2008년 발생한 강도 범죄 발생 건수를 나타낸 것이다. 강도 범죄 발생 건수를 분석하여 임계값의 범위와 범죄 발생 상태를 정의 하였다.

표 1. 범죄 발생 건수(강도)
Table 1.The incidence of crime (mugger)

	2006년	2007년	2008년
1월	352	326	263
2월	319	362	233
3월	380	377	436
4월	457	426	518
5월	724	503	420
6월	381	423	395
7월	344	311	401
8월	306	351	398
9월	357	401	512
10월	321	422	466
11월	383	261	389
12월	360	287	396

◆ 임계값 범위

S_1:0~305 S_2:351~500 S_3:501~650 S_4:651~800

◆ 범죄 발생 상태 (S)

$S = \{S_1, S_2, S_3, S_4\}$

식 (2)을 이용하여 2008년 8월~12월 까지 발생한 범죄 발생 건수를 이용하여 초기 확률 식(8)을 구한다.

◆ 범죄발생건수 : 398, 512, 466, 389, 396

S_2 S_3 S_2 S_2 S_2

◆ 초기 확률 : P (S_1:0, S_2:4 S_3:1, S_4:0)

P (0, 0.8, 0.2, 0) (8)

표 1의 범죄 발생 건수를 임계값의 범위에 매핑하여 상태들을 나타낸다.

$S_2\ S_1\ S_2\ S_2\ S_4\ S_2\ S_1\ S_1\ S_2\ S_1\ S_2\ S_2$

$S_1\ S_2\ S_2\ S_2\ S_3\ S_2\ S_1\ S_2\ S_2\ S_2\ S_1\ S_1$

$S_1\ S_1\ S_2\ S_3\ S_2\ S_2\ S_2\ S_2\ S_3\ S_2\ S_2\ S_2$

각 상태 $\{S_1, S_2, S_3, S_4\}$ 가 다른 상태로 전이하는 전이 횟수를 식(3)을 이용하여 전이확률로 만든 후, 전이행렬로 표현하면 식(9)로 표현된다.

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.29 & 0.52 & 0.14 & 0.05 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array} \quad (9)$$

그리고 식(9)의 상태전이 다이어그램은 그림 2와 같다.

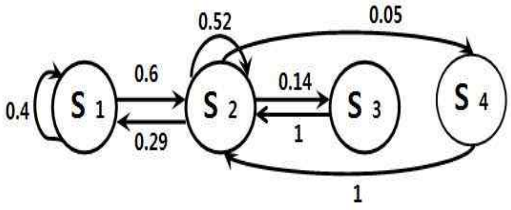

그림 2. 범죄에 대한 상태 다이어그램(강도)
Fig. 2. The state diagram for a crime (mugger)

초기 확률 식(8)과 전이행렬 식(9)을 구한 후 식(5)를 사용하여 범죄발생확률을 구할 수 있다.

$$(0 \quad 0.8 \quad 0.2 \quad 0)\begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.29 & 0.52 & 0.14 & 0.05 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad (10)$$

$$= (0.23 \quad 0.62 \quad 0.11 \quad 0.04)$$

구해진 식(10)에 의해 다음 달 강도 범죄 발생 확률은 상태가 S_2일 때 0.62로 가장 높다. 그러므로 다음 달에 강도 범죄발생건수는 S_2 상태인 351~500 사이에서 발생할 것으로 예측 된다. 본 논문은 다음 달에 발생할 범죄 발생 건수를 예측하기 위해 발생 확률 값에 평균값 식(6)과 최대 값 식(7)으로 나누어서 적용하고 그 값을 비교하였다.

◆ 평균 값 : 마지막 5개월간의 범죄 발생 건수 합의 평균값

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, M(S_i)$$

$$= 0.62 \times 432.2$$

$$= 267.96$$

◆ 최대 값: 2006~2008년 사이에 일어난 범죄 중 최고 발생 건수

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, Max(S_i)$$

$$= 0.62 \times 724$$

$$= 448.88$$

범죄발생건수를 예측한 결과 평균 값을 사용했을 때 약 268건, 최대 값을 사용했을 때 약 449 건이 발생되는 것으로 예측되었고, 실제 2009년 1월 달 강도 범죄발생건수는 484 건이었다.

1.3.2 범죄 유형 (강간)

표 2는 2006~2008년 발생한 강간 범죄 발생 건수를 나타낸 것으로, 범죄 발생 건수를 분석하여 임계값의 범위와 범죄 발생 상태를 정의 하였다.

표 2. 범죄 발생 건수(강간)
Table 2.The incidence of crime (rape)

	2006년	2007년	2008년
1월	806	1,011	850
2월	769	914	737
3월	1,029	1,034	930
4월	1,071	1,093	1,216
5월	1,317	1,330	1,376
6월	1,172	1,197	1,470
7월	1,248	1,315	1,617
8월	1,195	1,327	1,344
9월	1,311	1,081	1,695
10월	1,291	1,298	1,339
11월	1,381	1,082	1,230
12월	983	952	1,340

◆ 임계값 범위

S_1:0~750 S_2:751~1,100 S_3:1,101~1,450

S_4:1,451~1,800

◆ 범죄 발생 상태 (S)

$S = \{S_1, S_2, S_3, S_4\}$

앞의 절과 마찬가지로 식(2)를 이용하여 2008년 8월~12월 까지 발생한 범죄 발생 건수를 이용하여 초기 확률식(11)을 구한다.

◆범죄발생건수 : 1,334 1,695 1,339 1,230 1,340

S_3 S_4 S_3 S_3 S_3

◆ 초기 확률: P (S_1: 0, S_2: 0, S_3:4, S_4:1)

P (0, 0, 0.8, 0.2) (11)

표 2의 범죄 발생 건수를 임계값과 매핑하여 상태들을 나타낸 후, 각 상태 $\{S_1, S_2, S_3, S_4\}$가 다른 상태로 전이하는 전이횟수를 식(3)을 이용하여 전이확률로 만든 후, 전이행렬로 표현하면 식(12)으로 표현된다.

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.07 & 0.64 & 0.29 & 0 \\ 0 & 0.18 & 0.7 & 0.12 \\ 0 & 0 & 0.67 & 0.33 \end{pmatrix} \end{array} \quad (12)$$

초기 확률 식(11)과 전이행렬 식(12)를 구한 후 식(5)를 사용하여 범죄 발생 확률을 구할 수 있다.

$$(0 \quad 0 \quad 0.8 \quad 0.2)\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.07 & 0.64 & 0.29 & 0 \\ 0 & 0.18 & 0.7 & 0.12 \\ 0 & 0 & 0.67 & 0.33 \end{pmatrix} \quad (13)$$

$$= (0 \quad 0.14 \quad 0.69 \quad 0.16)$$

구해진 식(13)에 의해 다음 달 범죄 발생 확률은 상태가 S_3일 때 0.69로 가장 높다. 다음 달에 강간 범죄 발생 건수는 S_3 상태인 1,101~1,450 사이에서 발생할 것으로 예측 된다. 본 논문은 다음 달에 발생할 범죄 발생 건수를 예측하기 위해 발생 확률 값에 평균 값 식(6)과 최대 값 식(7)으로 나누어서 적용하고 그 값을 비교하였다.

◆ 평균 값 : 마지막 5개월간의 범죄 발생 건수 합의 평균값

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, M(S_i)$$
$$= 0.69 \times 1390$$
$$= 959.1$$

◆ 최대 값: 2006~2008년 사이에 일어난 범죄 중 최고 발생 건수

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, Max(S_i)$$
$$= 0.69 \times 1{,}695$$
$$= 1169.55$$

강간 범죄 발생 건수를 예측한 결과 평균 값을 사용했을 때 약 959건, 최대 값을 사용했을 때 약 1,170 건이 발생되는 것으로 예측되었고, 실제 2009년 1월 달 범죄발생건수는 876건이었다.

1.3.3 범죄 유형 (살인)

표 3은 2006~2008년 발생한 살인 범죄 발생 건수를 나타낸 것으로, 범죄 발생 건수를 분석하여 임계값의 범위와 범죄 발생 상태를 정의 하였다.

표 3. 범죄 발생 건수(살인)
Table 3.The incidence of crime(murder)

	2006년	2007년	2008년
1월	74	96	66
2월	69	83	50
3월	82	104	69
4월	81	89	105
5월	105	96	116
6월	96	91	92
7월	110	103	100
8월	114	98	116
9월	85	104	116
10월	92	108	105
11월	78	82	91
12월	78	70	94

◆ 임계값 범위

S_1:0~70 S_2:71~110 S_3:111~150

◆ 범죄 발생 상태 (S)

$S = \{S_1, S_2, S_3\}$

앞의 절과 마찬가지로 식(2)를 이용하여 2008년 8월~12월 까지 발생한 범죄 발생 건수를 이용하여 초기 확률 식(14)을 구한다.

◆범죄발생건수 : 116, 116, 105, 91, 94

$S_3 \quad S_3 \quad S_2 \quad S_2 \quad S_2$

◆ 초기 확률 : P (S_1: 0, S_2: 3, S_3:2)

P (0, 0.6, 0.4) (14)

표 3의 범죄 발생 건수를 임계값과 매핑하여 상태들을 나타낸 후, 각 상태 $\{S_1, S_2, S_3\}$가 다른 상태로 전이하는 전이 횟수를 식(3)을 이용하여 전이확률로 만든 후, 전이행렬로 표현하면 식(15)으로 표현된다.

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.08 & 0.81 & 0.11 \\ 0 & 0.75 & 0.25 \end{pmatrix} \end{array} \quad (15)$$

초기 확률 식(14)과 전이행렬 식(15)를 구한 후 식(5)를 사용하여 범죄(강간)발생 확률을 구할 수 있다.

$$(0 \;\; 0.6 \;\; 0.4)\begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.08 & 0.81 & 0.11 \\ 0 & 0.75 & 0.25 \end{pmatrix} \quad (16)$$

$$= (0.05 \;\; 0.79 \;\; 0.16)$$

구해진 식(16)에 의해 다음 달 범죄 발생 확률은 상태가 S_2 일 때 0.75로 가장 높다. 다음 달에 살인 범죄 발생 건수는 S_2 상태인 71~110 사이에서 발생할 것으로 예측 된다. 본 논문은 다음 달에 발생할 범죄 발생 건수를 예측하기 위해 발생 확률 값에 평균 값 식(6)과 최대 값 식(7)으로 나누어서 적용하고 그 값을 비교하였다.

◆ 평균 값 : 마지막 5개월간의 범죄(살인) 발생 건수 합의 평균값

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, M(S_i)$$

$$= 0.79 \times 104.4$$

$$= 82.476$$

◆ 최대 값: 2006~2008년 사이에 일어난 범죄(강간) 중 최고 발생 건수

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, Max(S_i)$$

$$= 0.79 \times 116$$

$$= 91.64$$

살인 발생 건수를 예측한 결과 평균값을 사용했을 때 약 83건, 최대 값을 사용했을 때 약 92건이 발생되는 것으로 예측되었고, 실제 2009년 1월 달 살인 발생건수는 79건이었다.

1.3.4 범죄 유형 (절도)

[표 4]는 2006~2008년 발생한 절도 범죄 발생 건수를 나타낸 것으로, 범죄 발생 건수를 분석하여 임계값의 범위와 범죄 발생 상태를 정의 하였다.

표 4. 범죄 발생 건수(절도)
Table 4.The incidence of crime(theft)

	2006년	2007년	2008년
1월	13,567	16,277	12,913
2월	12,544	16,056	12,047
3월	14,060	16,310	16,056
4월	15,412	17,759	20,527
5월	18,794	20,917	19,201
6월	16,696	21,130	20,200
7월	13,819	16,019	20,736
8월	13,835	15,714	16,518
9월	15,978	16,380	19,940
10월	17,762	21,152	19,873
11월	19,936	17,763	21,141
12월	18,342	17,053	24,112

◆ 임계값 범위

S_1 :0~13,000 S_2 :13,001~17,000 S_3 :17,001~21,000 S_4 :21,001~25,000

◆ 범죄 발생 상태 (S)

$S = \{S_1, S_2, S_3, S_4\}$

앞의 절과 마찬가지로 식(2)를 이용하여 2008년 8월~12월 까지 발생한 범죄 발생 건수를 이용하여 초기 확률 식(17)을 구한다.

◆ 범죄발생건수 : 16,518, 19,940, 19,873, 21,141, 24,112

S_2 S_3 S_3 S_4 S_4

◆ 초기 확률 : P (S_1: 0, S_2: 1, S_3:2, S_4:2)

P (0, 0.2, 0.4, 0.4) (17)

표 4의 범죄 발생 건수를 임계값과 매핑하여 상태들을 나타낸 후, 각 상태 $\{S_1, S_2, S_3, S_4\}$가 다른 상태로 전이하는 전이횟수를 식(3)을 이용하여 전이확률로 만든 후, 전이행렬로 표현하면 식(18)으로 표현된다.

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{pmatrix} 0.33 & 0.67 & 0 & 0 \\ 0.07 & 0.53 & 0.33 & 0.07 \\ 0.07 & 0.22 & 0.57 & 0.14 \\ 0 & 0.33 & 0.33 & 0.34 \end{pmatrix} \end{array} \quad (18)$$

초기 확률 식(17)과 전이행렬 식(18)를 구한 후 식(5)를 사용하여 범죄 발생 확률을 구할 수 있다.

$$(0 \quad 0.2 \quad 0.4 \quad 0.4)\begin{pmatrix} 0.33 & 0.67 & 0 & 0 \\ 0.07 & 0.53 & 0.33 & 0 \\ 0.07 & 0.22 & 0.57 & 0.14 \\ 0 & 0.33 & 0.33 & 0.34 \end{pmatrix} \quad (19)$$

$$= (0.04 \quad 0.33 \quad 0.43 \quad 0.21)$$

구해진 식(19)에 의해 다음 달 절도 발생 확률은 상태가 S_3 일 때 0.43으로 가장 높다. 다음 달 절도 범죄 발생 건수는 S_3 상태인 17,001~21,000 사이에서 발생할 것으로 예측 된다. 본 논문은 다음 달에 발생할 범죄 발생 건수를 예측하기 위해 발생 확률 값에 평균값 식(6)과 최대 값 식(7)으로 나누어서 적용하고 그 값을 비교하였다.

◆ 평균 값 : 마지막 5개월간의 범죄 발생 건수 합의 평균값

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, M(S_i)$$

$$= 0.43 \times 20{,}317$$

$$= 8736.31$$

◆ 최대 값: 2006~2008년 사이에 일어난 범죄 중 최고 발생 건수

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, Max(S_i)$$

$$= 0.43 \times 24{,}112$$

$$= 10{,}368.2$$

절도 발생 건수를 예측한 결과 평균값을 사용했을 때 약 8736건, 최대 값을 사용했을 때 약 10,368건이 발생되는 것으로 예측되었고, 실제 2009년 1월 달 살인발생건수는 14,305건이었다.

1.3.5 범죄 유형 (폭력)

표 5는 2006~2008년 발생한 절도 범죄 발생 건수를 나타낸 것으로, 범죄 발생 건수를 분석하여 임계값의 범위와 범죄 발생 상태를 정의 하였다.

표 5. 범죄 발생 건수(폭력)
Table 5.The incidence of crime(violence crime)

	2006년	2007년	2008년
1월	21,394	20,482	19,577
2월	19,262	16,711	15,321
3월	20,558	19,565	18,638
4월	19,208	20,671	20,815
5월	21,800	22,306	21,800
6월	21,098	22,199	23,718
7월	21,231	24,213	24,544
8월	20,558	23,312	21,979
9월	22,155	19,582	24,996
10월	22,621	24,609	24,506
11월	22,155	22,047	22,046
12월	19,150	19,561	23,559

◆ 임계값 범위

S_1:0~18,000 S_2:18,001~21,000 S_3:21,001~24,000

S_4:24,001~27,000

◆ 범죄 발생 상태 (S)

$S = \{S_1, S_2, S_3, S_4\}$

앞의 절과 마찬가지로 식(2)를 이용하여 2008년 8월~12월 까지 발생한 범죄 발생 건수를 이용하여 초기 확률 식(20)을 구한다.

◆범죄발생건수 : 21,979, 24,996, 24,506, 22,046, 23,559

S_3 S_4 S_4 S_3 S_3

◆ 초기 확률 : P (S_1: 0, S_2: 0, S_3:3, S_4:2)

P (0, 0, 0.6, 0.4) (20)

표 5의 범죄 발생 건수를 임계값과 매핑하여 상태들을 나타낸 후, 각 상태 $\{S_1, S_2, S_3, S_4\}$가 다른 상태로 전이하는 전이횟수를 식(3)을 이용하여 전이확률로 만든 후, 전이행렬로

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \end{array}\begin{array}{c} \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.15 & 0.46 & 0.31 & 0.08 \\ 0 & 0.33 & 0.47 & 0.2 \\ 0 & 0 & 0.8 & 0.2 \end{pmatrix} \end{array} \quad (21)$$

표현하면 식(21)으로 표현된다.

초기 확률 식(20)과 전이행렬 식(21)를 구한 후 식(5)를 사용하여 범죄 발생 확률을 구할 수 있다.

$$(0 \quad 0 \quad 0.6 \quad 0.4)\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.15 & 0.46 & 0.31 & 0.08 \\ 0 & 0.33 & 0.47 & 0.2 \\ 0 & 0 & 0.8 & 0.2 \end{pmatrix} \quad (22)$$

$$= (0 \quad 0.2 \quad 0.6 \quad 0.2)$$

구해진 식(22)에 의해 다음 달 폭력 발생 확률은 상태가 S_3 일 때 0.6으로 가장 높다. 다음 달에 폭력 발생 건수는 S_3 상태인 21,001~24,000 사이에서 발생할 것으로 예측 된다. 본 논문은 다음 달에 발생할 범죄 발생 건수를 예측하기 위해 발생 확률 값에 평균 값(식6)과 최대 값(식7)으로 나누어서 적용하고 그 값을 비교하였다.

◆ 평균 값 : 마지막 5개월간의 범죄 발생 건수 합의 평균값

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, M(S_i)$$
$$= 0.6 \times 23417$$
$$= 14{,}050.2$$

◆ 최대 값: 2006~2008년 사이에 일어난 범죄 중 최고 발생 건수

$$\text{범죄 발생 예측 건수} = \sum_{i=1}^{n} P(S_i)\, Max(S_i)$$
$$= 0.6 \times 24{,}996$$
$$= 14{,}997.6$$

폭력 발생 건수를 예측한 결과 평균값을 사용했을 때 약 14,050건, 최대 값을 사용했을 때 약 14,998건이 발생되는 것으로 예측되었고, 실제 2009년 1월 달 폭력 발생건수는 17,797건이었다.

2. 예측된 범죄 발생 건수 분석

강도, 강간, 살인, 절도, 폭력 5개의 범죄를 범죄 예측 모델링에 적용하여 범죄 발생 건수를 산출 하였다. 5개의 범죄에 평균 값과 최대 값을 적용한 결과 강도, 절도, 폭력의 경우 최대 값을 적용하였을 때, 강간, 살인의 경우 평균값을 적용하였을 때 예측 발생 건수와 실제 발생건수가 유사한 결과가 나왔다. 위의 결과는 표 6과 같다.

표 6. 범죄 발생 모델링 적용 값과 실제 범죄 발생 건수 비교
Table 6. Modelling crime and actual crime incidence compared to Apply

구분	예측값		실제 발생 건수	차이	
	평균값	최대값		평균값	최대값
강도	268	449	484	216	35
강간	959	1,170	876	83	294
살인	83	92	79	4	13
절도	8,736	10,368	14,305	5,569	3,937
폭력	14,050	14,998	17,797	3,747	2,799

IV. 결 론

본 논문은 미래 예측 모델인 마코프 체인을 이용한 범죄 예측 모델링을 제시하였고, 5개의 범죄를 예측해 보았다. 본 논문은 마코프 체인을 이용한 이전 연구와는 다르게 연구 대상 기간의 범죄 평균값 외에 최대값을 넣어서 미래에 발생할 범죄 발생 건수를 예측해 보았다. 그 결과 중간 값의 선택에 따라 최근 발생한 범죄발생건수의 평균값을 사용하느냐, 범죄 발생 최대 값을 적용하나에 따라 예측 값이 변하였다. 3개의 범죄(강도, 절도, 폭력)에서는 최대 값을 넣었을 때 실제 발생한 범죄 발생건수와 유사하였고, 2개의 범죄(강간, 살인)에서는 평균값을 넣었을 때 실제 발생 건수와 유사하였다. 이 결과 범죄 예측에 있어서는 중간 값으로 평균값을 선택하는 것보다 최대 값을 넣는 것이 더 실제 발생 건수를 예측하는데 도움이 될 것으로 예상된다. 본 논문에서 제시한 마코프 체인을 이용하면 범죄 예측 모델링을 적용하면 범죄가 발생할 발생 건수를 예측 할 수 있으므로, 범죄에 대한 예방 정책을 수립할 때 도움을 줄 수 있다. 그러나 마코프 체인으로 나온 결과는 가까운 미래는 예측 할 수 있으나 먼 미래는 예측 할 수 없고, 연쇄 살인범과 같은 특수한 범죄의 경우 예측하기 어려운 점이 있다. 향후에는 이러한 단점을 보안한 연구가 진행되어야 할 것이다.

참고문헌

[1] Brown, M.A., "Modelling the Spatial Distribution of Suburban Crime," Economy Geography, Vol. 58, No3, pp. 247-261, July 1982.

[2] Kamber, T., Mollenkopt, H., and Ross, A. "Crime, Space, and Place : An Analysis of Crime Patterns in Brooklyn", inn Goldsmith V., Mguire G., Mollenkopf, H. and Ross, A.(eds.), Analyzing Crime Patterns: Frontiers of Practice Sage, pp121-136, 2000.

[3] Lee Sang-Hyun, "Evaluation of Crime Prevention Performance of Urban Spatial Structure through Network Analysis", Journal of the architecture institute of Korea planning & design, Vol27, No 8, pp. 243-250, 8, 2011.

[4] Dong-Suk Hong, Joung-Joon Kim , Hong-Koo

Kang , Ki-Young Lee, Jong-Soo Seo, Ki-Joon Han, “Base Location Prediction Algorithm of Serial Crimes based on the Spatio-Temporal Analysis”, Journal of Korea spatial information society, Vol10, No 2, pp. 63-79, 6. 2008.

[5] Park Cheol-Hyun, “Specialization in Criminal Career : Markov-Chain Analysis”, Korean Criminological Review, pp. 243-273, 3, 2003.

[6] Charles M. Grinstead, “Introduction to Probability: Second Revised Edition”, American Mathematical Society, pp405-406, 1997.

[7] Won-Hyung Park, Young-Jin Kim, Dong-Hwi Lee, Kui-Nam J Kim, “A Study on Prediction of Mass SQL Injection Worm Propagation Using The Markov Chain”, Journal of the Korea Institute of Information Security and Cryptology, Vol 8, No4, pp.174-181, 12. 2008.

[8] Young-Gab Kim, Young-kyo Baek, Hoh Peter In, Doo-Kwon Baik, “A Probabilistic Model of Damage Propagation based on the Markov Process”, Journal of KIISE, Vol33, No8, pp.524-535, 8. 2006.

[9] Seung-Hun Lee, Byeong-Sup Moon, Bum-Jin Park , “The Bus Delay Time Prediction Using Markov Chain”, The Journal of The Korea Institute of Intelligent Transport Systems, pp.1-10, 6. 2009.

[10] Kim Young-Jin, Park Cheol-Soo, “Prediction of Occupant″s Presence in Residential Apartment Buildings using Markov Chain” Korea Institute of Architectural Sustainable Environment and building System. 2008 autumn conference, pp116~121, 2008.

[11] Hee Tae-Lee, Jae-Chul Kim, “A Study on The Prediction of Number of Failures using Markov Chain and Fault Data”, KIIEE Annual Autumn Conference 2008, pp. 363-366, 10.2008

저 자 소 개

정 영 석

1998: 서남대학교 물리학과 이학사.
2000: 배재대하교 물리학과 이학석사.
2009: 공주대학교 멀티미디어공학과 공학석사.
현 재: 공주대학교 컴퓨터 공학과 박사수료
관심분야: 시뮬레이션, 클라우드 컴퓨팅, 영상처리, 정보보안
Email : merope@kongju.ac.kr

김 진 묵

2006: 광운대학교 전자계산학과 공학박사.
2006: 선문대학교 컴퓨터과학과 연구교수.
현 재: 선문대학교 IT교육학부 조교수
관심분야: 정보보안, 센서네트워크 보안, 클라우드 컴퓨팅 보안
Email : calf0425@sunmoon.ac.kr

박 구 락

1986: 중앙대학교 전기공학과 공학사.
1988: 숭실대학교 전자계산학과 공학석사.
2000: 경기대학교 전자계산학과 이학박사.
현 재: 공주대학교 컴퓨터공학부 교수
관심분야: 정보경영, 정보통신, 전자상거래
Email : ecgrpark@kongju.ac.kr