

# 시험 보고서

발 행 번 호 : 2015 - 032A

회 사 명 : 전자부품연구원(KETI)

시 료 명 : 범죄 패턴 예측 알고리즘

시 험 명 : 범죄 패턴 예측 정확도 평가

상기 사항은 **CEWIT KOREA**에서 입회하에 실시한 성적서임을 증명함.



본 시험성적서는 의뢰자가 제공한 시료의 시험 결과로써 시료명은 의뢰자가 제시한 것입니다. 본 시험 및 분석은 단지 제품의 품질 및 신뢰성 향상을 위한 시험 분석입니다. 이 성적서는 용도 이외의 사용을 금하며, 의뢰자가 임의로 사용하여 발생하는 문제에 대하여 책임지지 않습니다.

119 Songdo Moonwha-ro, Incheonl, Korea 406-840

Tel : +82 32 626 1533



Fax : +82 32 626 1559

# 시 험 결 과

## 1. 시험 고객

회 사 명	전자부품연구원 (KETI)		
대 표 자	박 청 원		
주 소	경기도 성남시 분당구 새나리로 25		
신 청 자	박 용 주	접 수 일	2015/08/17
시 료 명	범죄 패턴 예측 알고리즘	시 료 수	-
시험성적서 용도	범죄 패턴 예측 알고리즘 성능 평가		
총 페이지 수	6 페이지		

## 2. 시험조건 및 시험결과

시험 항목	범죄 패턴 예측 정확도 평가 (TP rate 60% 달성 여부 확인)			
시험 방법	시험의뢰자 제공 실행 방법에 따라 테스트 데이터에 적용			
시험 결과	Pass. 예측 정확도 72% 이상 (시험 결과 및 별첨 참조)			
시험 기간	2015/08/17 ~ 2015/08/19			
시험 환경	H/W	i5-560,8G,GT420M	OS	Windows 7 64bit
시 험 자		김 기 영  (sign)		
승 인 자		이 근 명  (sign)		

## 세 부 시 험 내 역

- 목적** : 범죄 패턴 예측 알고리즘을 실제 범죄 발생 CCTV 영상에서 행동 패턴을 추출한 데이터에 적용하여 범죄 유형 및 범죄발생 여부 예측 정확도가 목표를 만족시키는지 여부를 검증
- 시료** : 전체 범죄 패턴 예측 시스템 구성은 아래 그림과 같다. 그 중 테스트 대상이 되는 부분은 아래 그림의 “범죄 유형 및 범죄발생 여부 예측”을 수행하는 두 번째 모듈로 CCTV 스트림에서 첫 번째 모듈을 거치면서 가공된 다차원 행동 패턴 특성 변수 (장소, 시간, 행동특성 등의 nomial variable)을 입력받아 다섯 개의 범죄 클래스 (폭행, 절도, 강도, 성폭행, 살인) 중 한 유형에 속하는지 여부를 예측한다. 실행 프로그램은 txt 형식으로 행동 특성 변수를 입력받아 역시 txt 형식으로 예측 결과를 출력한다.



- 데이터** : 의뢰업체에서 제공한 실제 범죄 발생 CCTV 영상에서 추출한 행동 패턴 데이터. 총 1589개의 case가 있으며, case 별로 각 한 개의 ID와 22개의 변수가 있다. (순서대로 ID, 성별, 나이대, 요일, 범죄발생장소, 범행시간대, 가해자음주여부, 피해자음주여부, 접근미행여부, 범행도구, 가택침입여부, 폭행특성, 납치여부, 기다렸다달치기, 길을걷다폭행, 심신미약의피해자도와중, 도구로열려고함, 시간대, 피해자성별, 고함욕설, 범죄장소특징, 목살잡음, 몸싸움여부) 테스트 결과의 경우, ID에 이어서, 각 범죄 패턴 예측 확률이 실수로 나열되어 있다. (폭행, 절도, 강도, 성폭행, 살인 순)

입력:	1189	0	1	4	10	1	1	1	0	1	0	0	0	0	0	0	0	1	0	2	0	0															
출력:	1189	모델 1	0.007	0.988	0.004	7E-04	4E-04																														
답안:	1189	0	1	0	0	0																															
테스트 데이터의 입력/출력/답안 케이스 예시																																					

- 샘플링** : 범죄 패턴 예측 알고리즘은 기본적으로 케이스 별로 결과를 산출한다. 테스트 셋 구성은 1차로 전체 테스트 케이스 (1589건)에 대하여 테스트를 진행하고, 2차로 각

범죄 유형별로 테스트 케이스를 구성해서 테스트를 진행하고, 3차로 전체 테스트 케이스를 10개의 임의의 subset으로 나누어 각 세트에서의 정확도를 측정하였다. (5회 반복)

## 5. 시험 조건

가. 성능 지표 정의와 성능 목표치 근거

- (성능 지표) 범죄 패턴 예측률 정확도 평가 (정확도(true-positive rate) 60% 이상)
- 범죄 패턴 예측률 : 가공된 정량적 범죄인자(범죄 유형별 가중치 적용)를 엔트로피 엔진에 입력으로 인가하였을때 나오는 범죄유형 선택의 확률적 정확도
- (성능 목표치 근거) 범죄 패턴인식의 정확성은 엔트로피 필터링 엔진의 고차원 Depth에 의해 결정되며, 유사한 특성을 갖는 예측 시스템(불량률 예측 모델 등등,,) 적용시 약 80%(최종목표)의 정확성을 보임

## 6. 시험 방법

가. 1차 : 전체 데이터 (1589건)을 하나의 테스트셋(testInput1.txt)으로 구성하여 프로그램을 수행하고 성능을 측정. 프로그램은 txt 형식 (탭 구분)의 입력을 받아 예측 결과를 txt 형식 (탭 구분)의 출력파일에 기록한다. 예측 결과를 엑셀을 이용하여 ground truth와 비교하여 TP rate을 구하였다.

```
> CrPrediction4Test.exe testInput1.txt testOutput1.txt ↵
```

### 프로그램 수행 방법

나. 2차 : 데이터를 각 범죄 유형별로 미리 분류하여 각 유형별로 예측 정확도를 구하였다. 각 범죄 유형별 테스트 케이스 개수는 폭행 620건, 절도 733건, 강도 74건, 성폭행 143건, 살인 19건이며, 각 유형별로 testInput2.txt, testInput3.txt, testInput4.txt, testInput5.txt, testInput6.txt로 입력 파일을 구성하여 총 5번에 걸쳐서 프로그램을 실행하였다.

다. 3차 : 전체 테스트 케이스를 10개의 랜덤 subset으로 나누어 각 세트 별 정확도를 측정하였다. 엑셀에서 rand() 함수를 통해 임의수를 각 케이스에 부여하고, 그 임의의 수로 sorting을 하여 랜덤 subset을 생성하였다. 각 세트는 159~158개의 케이스로 구성되고, 각 세트마다 프로그램을 수행하여 예측 정확도를 측정하였다. 랜덤하게 subset을 구성하기 때문에 실험 신뢰도를 높이기 위해 이를 총 5회 반복하였다. (10세트 \* 5회 = 총 50회 수행)

## 7. 시험 결과

가. 1차 : 전체 1589건에 대해 범죄 패턴 예측 프로그램을 수행한 결과, 1273건에 대해 올바른 예측을 하였다. **정확도는 80.1%로 성능 지표를 만족시켰다.**

나. 2차 : 각 범죄 유형별 테스트 수행 결과 예측 정확도는 다음과 같다. 폭행 620건 중 552건을 맞췄 (89.0%), 절도 733건 중 699건을 맞췄 (95.4%), 강도 74건 중 22건을

맞춤 (30.0%), 성폭행 143건 중 0건을 맞춤 (0.0%), 살인 19건 중 0건을 맞춤 (0.0%). 결과를 보면 테스트 케이스 수가 적을수록 현격하게 성능이 떨어짐을 알 수 있는데 이해 대해서는 부록에서 분석을 수행하였다.

다. 3차 : 전체 테스트 케이스를 10개의 랜덤 서브셋으로 나누어 각각의 세트에 패턴 예측 프로그램을 수행하기를 5회 반복하였다. 각 경우에 예측 정확도는 아래 테이블과 같다. 아래 표를 보면 모든 경우에 **최소 72% 이상의 예측 정확도를 보여 성능 지표를 만족함**을 알 수 있다.

시 행	1	2	3	4	5	6	7	8	9	10
시 행1	75.5%	80.0%	75.5%	79.9%	81.85	84.9%	82.4%	82.4%	84.9%	74.0%
시 행2	85.5%	77.4%	77.4%	73.6%	82.4%	79.2%	73.0%	85.5%	83.6%	83.5%
시 행3	74.8%	82.4%	84.3%	81.8%	78.6%	79.2%	80.5%	81.8%	80.5%	77.2%
시 행4	86.2%	73.0%	74.8%	83.0%	83.0%	79.2%	77.4%	79.2%	84.9%	80.4%
시 행5	79.9%	76.7%	82.4%	84.9%	74.2%	72.3%	78.6%	86.8%	79.2%	86.1%

**전체 테스트 케이스를 10개의 랜덤 서브셋으로 나누어 각각의 예측 정확도를 측정한 결과**

## 8. 부록

### 가. 실험 데이터

- 실험 데이터는 별도 파일로 전달하며, 파일명은 아래와 같이 구성되어 있다.

<b>1차 실험</b>	
실험 입력 데이터	testInput1.txt
실험 결과 데이터	testOutput1.txt
1차 실험 및 2차 실험 예측 정확도 분석	testOutput-t0.xlsx
<b>2차 실험</b>	
실험 입력 데이터(폭행, 절도, 강도, 성폭행, 살인)	testInput2.txt ~ testInput6.txt
실험 출력 데이터(폭행, 절도, 강도, 성폭행, 살인)	testOutput2.txt ~ testOutput6.txt
<b>3차 실험</b>	
실험 입력 데이터(1차)	testInput-r1.txt ~ testInput-r10.txt
실험 입력 데이터(2차)	testInput-r20.txt ~ testInput-r29.txt
실험 입력 데이터(3차)	testInput-r30.txt ~ testInput-r39.txt
실험 입력 데이터(4차)	testInput-r40.txt ~ testInput-r49.txt
실험 입력 데이터(5차)	testInput-r50.txt ~ testInput-r59.txt
실험 출력 데이터(1차)	testOutput-r1.txt ~ testOutput-r10.txt
실험 출력 데이터(2차)	testOutput-r20.txt ~ testOutput-r29.txt
실험 출력 데이터(3차)	testOutput-r30.txt ~ testOutput-r39.txt
실험 출력 데이터(4차)	testOutput-r40.txt ~ testOutput-r49.txt
실험 출력 데이터(5차)	testOutput-r50.txt ~ testOutput-r59.txt
3차 실험 예측 정확도 분석(1차~5차)	testOutput-t1.xlsx ~ testOutput-t5.xlsx
<b>기타 실험</b>	
Weka 입력 데이터	분석_데이터.csv
Weka 분석 결과(SVM,MLP,RF)	weka_svm.txt, weka_mlp.txt, weka_rf.txt
<b>실험 데이터의 파일명 목록</b>	

### 나. 추가 검토 의견

- **(데이터셋 개선)** 모집단을 CCTV 영상 전체라고 가정할 경우 CCTV 영상 중 범죄가 일어나지 않은 경우가 대부분일 것을 감안하면 학습 데이터에 TN example을 포함시킬 필요가 있음. 특히, 전체 테스트 데이터 중 319건은 시간대, 장소특성 등을 제외하고 모든 범죄 행동과 관련된 variable이 negative 임에도 범죄 상황이어서 현재로서는 TN와 구분하기 어려울 것으로 예상됨
- **(성능 지표 개선)** 또한, 테스트셋에 negative example을 포함시킬 경우, 성능 지표에 있어서 Precision 및 Recall을 모두 감안해야 할 것으로 보임 (f-measure와 같이)
- **(알고리즘 개선 방안)** 강도, 성폭행, 살인과 같은 위험도는 높으나 빈도가 낮은 범죄에 대한 데이터가 부족하여 예측 정확도가 현저히 떨어지는 문제가 발생함. 실제로 Weka 툴킷을 이용해 random forest를 적용하여 10-fold cross-validation으로 학습을 수행한 경우, 강도와 성폭행 측면에서는 본 알고리즘 대비 다소 높은 성능을 보였음. 기본적으로 빈도가 낮은 범죄에 대한 데이터 확보가 우선 필요하나, 중대 범죄에 대해서만 별도 학습하여 decision tree로 통합한다던지 하는 방법 등을 적용하여 중대 범죄에 대한 예측 정확도를 높일 필요가 있음

- End Page