대한수사과학지 제1권 제2호
Journal of Forensic and Investigative Science Vol. 1, No. 2 (2006) p. 5～19

데이터 마이닝의 범죄수사 적용 가능성

김준우*, 손중권**, 이상한
경찰청 수사과*, 경북대학교 자연과학대학 통계학과**, 경북대학교 의과대학 법의학교실

Usefulness of Data Mining in Criminal Investigation

Joon Woo Kim*, Joong Kweon Sohn, Sang Han Lee**
Criminal Investigation Division, National Police Agency, Department of Statistics, Kyungpook National University**, Department of Forensic Medicine, Kyungpook National University School of Medicine*

Abstract - Data mining is an information extraction activity to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis. Law enforcement agencies deal with mass data to investigate the crime and its amount is increasing due to the development of processing the data by using computer. Now new challenge to discover knowledge in that data is confronted to us. It can be applied in criminal investigation to find offenders by analysis of complex and relational data structures and free texts using their criminal records or statement texts. This study was aimed to evaluate possibile application of data mining and its limitation in practical criminal investigation.

Clustering of the criminal cases will be possible in habitual crimes such as fraud and burglary when using data mining to identify the crime pattern. Neural network modelling, one of tools in data mining, can be applied to differentiating suspect's photograph or handwriting with that of convict or criminal profiling. A case study of in practical insurance fraud showed that data mining was useful in organized crimes such as gang, terrorism and money laundering. But the products of data mining in criminal investigation should be cautious for evaluating because data mining just offer a clue instead of conclusion. The legal regulation is needed to control the abuse of law enforcement agencies and to protect personal privacy or human rights.

Keywords : data mining, organized crimes, criminal investigation

Corresponding author : Sang Han Lee, M.D., Ph.D.
Tel : 053-420-4885, Fax : 053-422-4712
sanghan1@knu.ac.kr

Ⅰ. 서 론

수사란 범죄혐의가 있는 경우에 그 혐의 사실을 해명하기 위하여 범인을 발견, 보전하며 증거를 수집, 확보하는 활동을 말한다. 따라서 수사관은 범죄자를 특정하고, 혐의를 입증하기 위해 많은 사람들로부터 정보를 얻고, 계좌거래 내역이나 통화 내역 등과 같은 다양한 자료를 수집하게 된다. 사회 전반에 걸쳐 컴퓨터에 의한 정보처리가 이루어지면서 범죄 수사를 위해 수집되는 그 자료의 양 또한 기하급수적으로 늘어나고 있다. 그러나 이러한 정보가 담고 있는 의미와 각 정보의 연결고리를 밝혀내지 못한다면 과거 발생된 범죄를 입증할 수 없을 뿐더러 현재 일어나는 범죄 현상을 설명할 수 없고, 앞으로 일어날 범죄를 예측할 수도 없다. 또한 이와 같은 수사 활동으로 수집된 방대한 양의 정보를 보관하고 있다 하더라도 한 사건에 있어 적재적소에 활용할 수 없다면 무용지물이 되고 말 것이다. 따라서 우리는 이러한 많은 양의 정보와 자료를 처리하기 위해 이를 데이터화하여 관리를 하고 있는데, 이것이 우리가 흔히 부르는 데이터베이스이며, 현재 경찰 역시 범죄자와 범죄 자료에 대한 데이터베이스를 구축하여 관리하고 있다.

한편, 한 사건에 관련된 자료의 양이 늘어나면 늘어날수록, 범죄와 범죄 자료에 대한 데이터가 축적되면 축적될수록 이의 활용은 더욱 더 난해해지는 새로운 문제에 직면하게 되었다. 데이터가 넘쳐나는 상황에서 정작 특정 사건이나 특정인의 수사에 필요한 관련 정보를 추출해내는 것은 수사관 개개인의 이해력을 넘어서고 있으며, 어떤 데이터들은 접근조차 되지 않고 있는 것이다. 따라서 많은 양의 다양한 데이터베이스를 구축함과 동시에 이러한 데이터의 옥석을 구별하고, 흩어진 데이터 속에 숨겨진 의미는 무엇인지, 사건 수사에 필요한 데이터는 어떤 것이 있으며 이를 어떻게 활용하여야 하는지에 대한 새로운 방법론을 모색해야 할 필요성이 있다.

수집된 정보나 구축된 데이터베이스에서 의미 있는 정보를 가려내는 기법 중 하나가 바로 데이터 마이닝(data mining)이다[1]. 데이터 마이닝은 대량의 데이터로부터 지식을 추출하는 과정 또는 기법을 말하는 것으로, 쌓여가는 데이터 속에서 의사결정에 필요한 중요한 정보를 파악하고, 각 데이터간의 패턴을 인식하는 문제에 대한 대안인 것이다[2]. 이러한 데이터 마이닝은 많은 자료를 접하게 되는 범죄 수사에 있어서도 그 활용 가능성이 있으나 아직 이에 대한 연구는 미미한 편이다.

본 연구는 각 사건마다 접하는 데이터의 종류가 다양해지고, 그 양이 많아지고 있는 상황에서 이를 분석하는 새로운 방법론이 필요하다는 인식하에 최근 데이터 처리의 대안으로 대두되고 있는 데이터 마이닝의 수사 적용 가능성에 대해 고찰하고자 한다. 구체적으로 본 연구에서 다루고자 하는 문제는 다음과 같다.

첫째, 데이터 마이닝의 정의와 그 기법을 개괄적으로 살피고, 각 기법이나 모델의 유형들이 어떤 문제에 효과적으로 적용될 수 있는지를 비교, 분석한다.

둘째, 데이터 마이닝 기법이 실제 수사에 있어서 어떻게 적용될 수 있는지 그 가능성에 대해 살펴보고, 각 기법들이 적용 가능한 범죄유형과 수사기법을 살펴본다.

셋째, 데이터 마이닝의 범죄수사 적용에 있어 효용성 및 한계를 살펴보고 적용 타당성을 검토하여 활용방안을 제시하고자 한다.

Ⅱ. 연구 방법과 연구 내용

데이터 마이닝의 수사 적용 가능성에 대해 본 연구에서는 크게 두 가지 방법으로 접근하고자 한다. 우선, 데이터 마이닝에 대한 개념과 기능을 정리하고, 이의 적용과정을 파악하기 위해 문헌 연구를 실시하였다. 다양한 학문적 접근 방법인 데이터 마이닝의 정의와 절차를 검토하고, 그 목적과 구체적인 기법을 살펴봄으로써 실제 수사에 적용 가능한 방법론을 모색하였으며, 외국 선행 연구 및 적용사례를 소개하였다.

다음으로, 구체적 사안에 대해 데이터 마이닝이 어떻게 적용될 수 있는가를 파악하기 위해 사례 분석을 실시하였다. 특정 단서를 바탕으로 조직적 체계를 파악하기 위해 본 연구에서는 보험사기 혐의 용의자에 대한 시각화 분석을 적용하였다. 본 연구에 사용된 자료는 실제 사건 접수된 것으로 금융감독원의 자료를 바탕으로 하였다. 인적자료에 관한 기재는 생략하였고, 특정 회사의 명칭도 가상으로 하였다. 사례 분석에서 이용된 보험회사 자료의 데이터 자체를 직접 확보하는 것은 개인 정보에 관한 것이고, 각 관련자와 보험 회사의 수가 다수인 관계로 어려운 점이 있어 테이터 마이닝 툴(tool)이나 시각화 툴을 이용할 수 없

었고, 따라서 미리 확보된 자료를 바탕으로 데이터 마이닝 과정을 수작업 해둔 것임을 미리 밝혀둔다. 마지막으로 위에서 살펴본 과정을 통해 수사 적용에 있어 데이터 마이닝이 가지는 효용성과 한계점은 무엇인지 정리하고, 이에 따른 앞으로의 전망과 연구 과제를 제시하였다.

Ⅲ. 데이터 마이닝의 이해

1. 데이터 마이닝의 정의 및 유사개념

데이터 마이닝은 일반적으로 대량의 데이터 내부로부터 유도된 새로운 모델로 과거에는 알지 못했던 실행 가능하고, 의미 있는 정보나 지식을 추출하는 일련의 과정을 말한다. 쉽게 말해 데이터 속에 함축된 유용한 지식이나 패턴을 찾아내는 기술이다. 최근 20년간 컴퓨터 환경이 발전하고, 데이터베이스 기술이 축적됨에 따라 우리는 많은 양의 데이터 속에서 무엇을 찾아내고, 이용할 수 있는지를 알 수 없었다. 이와 같은 데이터 홍수의 시대에 규칙을 추론함으로써 의사 결정을 지원하고 그 효과를 예측하려는 자료 분석 및 모형 선정 과정이 바로 데이터 마이닝인 것이다[1].

이러한 데이터 마이닝은 새로운 기술이라기보다는 기계-학습 분야(machine learning)나 인공지능(Artificial Intelligence, AI) 분야, 통계학, 데이터베이스 기술, 시각화(visualization) 기술 등이 모두 기여를 한 학제적 연구 분야로, 지식 마이닝(knowledge mining from database), 지식 추출(knowledge extraction), 데이터/패턴 분석(data/pattern analysis), 지식발견(Knowledge Discovery in Database, 이하 KDD)과 혼용되고 있으며, 많은 사람들은 데이터 마이닝을 지식발견(KDD)과 동의어로 취급하고 있다[2].

그러나 1995년 몬트리얼에서 열린 제1회 국제 KDD/Data mining 학술대회에서 “KDD는 데이터로부터 지식을 추출하는 전 과정”을 설명하는 뜻으로 해석하고, “데이터 마이닝은 KDD 과정에서 탐사 단계만을 뜻하는 의미로 사용한다.”고 제안하였고[3], 또한 후술할 데이터베이스와 질의(structural query language, 이하 SQL), 데이터 웨어하우스(data warehouse)와 OLAP(on-line analytical processing)의 상대적인 개념 차이를 감안한다면 데이터 마이닝을 지식발견(KDD) 과정의 한 단계로 파악하는 것이 타당하다고 생각된다. 아직 데이터 마이닝과 관련하여 명확한 개념의 정의가 이루어진 것은 아니며, 또 다른 학자들의 경우 좁의 의미로 데이터 마이닝의 개념을 한정시킨다고 할지라도 실제 데이터 마이닝을 위해서는 데이터 준비 및 전처리 과정이 필수 불가결한 것이므로 결국 데이터 마이닝 과정이 지식발견(KDD) 과정과 동일하다는 이유로 데이터 마이닝 자체를 지식발견(KDD)와 동일한 개념으로 받아들이기도 한다. 이는 데이터 마이닝을 넓은 개념으로 파악하는 것이다. 따라서 본 논문에서는 위 몬트리얼 학술대회에서 밝힌 바와 같이 데이터 마이닝을 지식탐사단계 중 패턴 및 지식 추출을 위한 탐사단계로서의 좁은 의미로 파악하도록 한다.

2. 데이터 마이닝과 기존 데이터 처리 기법과의 비교

데이터 마이닝은 자동화되고 지능을 갖춘 데이터베이스 분석기법이기는 하지만 기존의 분석 기법을 대체하는 것이 아니라 보완하는 기능을 하고 있다. 그러나 데이터 마이닝을 통해 얻을 수 있는 추가적인 정보의 가치는 그 활용도에 따라 무한하다고 할 수 있는데, 이를 살펴보면 다음과 같다.

예를 들어 유통업체가 수년간의 영업활동을 통하여 상당한 양의 고객 구매데이터를 수집하여 관리를 하고 있다고 가정할 때 데이터베이스의 가장 기초적인 일반적인 질의(query; 통상 SQL(structured query language)로 대표되는데, SQL은 관계형 데이터베이스의 조작과 관리에 사용하는 데이터베이스 하부 언어로서 입력 테이블로부터 원하는 출력 테이블을 사상시키는 언어이다)를 통해 특정 시점에서 특정 지역의 특정 상품에 대한 판매 현황에 대한 테이블 형태의 결과를 얻어낼 수 있다. 그러나 이러한 데이터베이스의 질의를 통해서는 특정 지역에서 어떤 상품들이 팔렸는지 월별 및 고객 연령대별로 분석하는 복합적인 결론을 내리기 힘들고, 이를 위해서는 숙련된 사용자에 의하여 적당한 가설 하에 수많은 질의를 통해서만 가능하였다.

그러나 이러한 문제를 해결하고, 사용자의 의사결정을 지원하기 위해 많은 양의 이질(異質) 데이터를 사용자 관점에서 주제별로 통합시킨 데이터 웨어하우스를 통한 OLAP(통상 SQL(structured query language)로 대표되는데, SQL은 관계형 데이터베이스의 조작과 관리에 사용하는 데이터베이스 하부 언어로서 입력 테이블로부터 원하는 출력 테이블을 사상시키는 언어이다) 기법이 등장하게 되

었다. OLAP는 다차원 정보에 직접 접근하여 대화식으로 정보를 분석하고 의사결정에 활용하는 기법이라고 할 수 있는데, 이를 통하여 위 예에서 제시한 문제의 지역, 상품, 기간, 연령이라는 4개의 독립된 속성으로 구성된 4차원 분석이 가능하게 된 것이다[2].

이러한 기존의 분석 기법들은 기준이 되는 속성들을 사전에 파악할 수 없기 때문에 원하는 결과를 얻기 위해서는 수많은 시도와 실패가 요구된다. 즉, 앞서 살펴본 질의나 OLAP와 같은 기존의 데이터 분석 도구들은 사용자가 자신의 경험에 비추어 가설을 세우고 이 가설을 바탕으로 데이터 접근을 시도한 후 결과를 검토하여 가설의 타당성을 확인할 뿐이다. 그러나 이러한 기법들은 아무리 잘 구축한다고 하더라도 적절한 가설을 세울 수 없는 상황에서는 적용하기 곤란하며, 가설을 검증하는 것 이외에는 유용한 정보를 생성할 수 없다. 그러나 데이터 마이닝은 많은 양의 데이터로부터 짧은 시간 내에 숨겨진 의미와 패턴을 인식함으로써 유용한 가설을 추출하여 새로운 정보를 발견하게 되는데, 이것이 데이터 마이닝이 기존의 다른 기법과 차별되는 점이라 하겠다[3].

위에서 제시한 수년간의 고객 구매데이터에서 본다면 어떤 제품을 구매하는 고객들의 특성이나 특정 상품의 매출을 높이기 위한 전략을 도출해낼 수 있는 것이다. 그러나 여기서 우리가 유념해야 할 것은 데이터 마이닝은 결코 독립되어 존재할 수 없다는 점이다. 즉 많은 양의 데이터 중에서 기존의 기법으로 얻을 수 없는 추가적인 정보만을 제공한다는 것이다. 따라서 기존 기법을 적용할 수 없는 환경에서 데이터 마이닝을 사용하려는 것은 데이터가 가지고 있는 상당량의 기본 정보를 무시하고, 극히 미약한 숨겨진 정보만을 찾겠다는 발생과 같다. 이는 곧 데이터 마이닝이라는 것은 기존 기법을 적용할 수 있는 데이터베이스나 데이터 웨어하우스의 구축 없이는 적용 불가능하다는 것을 의미하는 것이며, 또한 기존의 접근 방법으로 파악 가능한 사항을 데이터 마이닝을 통하여 찾으려는 실수를 범하지 말아야 한다는 것을 의미하기도 한다.

3. 지식 발견 과정(KDD processing)과 데이터 마이닝

지식 발견 과정은 사용자와 상호 작용을 하면서, 여러 단계로 구성된 반복적인 업무 수행을 통해 점진적으로 지식을 발견하는 모형화 과정이다. 각 학자들은 이러한 지식 발견 절차와 관련하여 조금씩 다른 틀을 마련하고 있지만 가장 널리 인용되고 있는 Fayyad 등의 개념을 인용하여 설명하도록 한다[4]. 즉, 데이터베이스로부터 추출된 원자료(raw data)로부터 유용한 정보를 얻기 위해서는 데이터 선택, 데이터 전처리, 데이터 변환, 데이터 마이닝, 해석 및 평가의 다섯 단계를 거치게 된다.

1) 데이터 선택(data selection, sampling)

지식 발견 과정의 가장 첫 번째 단계로서 발견하고자 하는 지식에 대한 목표를 설정한 후 분석할 데이터를 선정하고 수집하여 대상 데이터 셋을 만들고 구체적 분석 알고리즘이 적용될 변수 집합이나 데이터 표본을 추출, 정리하는 단계이다. 현재 보유하고 있는 데이터 중에서 관심의 대상이 되는 클래스를 선정하는 과정인 것이다.

2) 데이터 전처리(data preprocess)

우리가 분석하고자 하는 데이터는 실제 생활을 반영한 것이기 때문에 불완전하고(관심있는 속성 값이 없거나 집계 데이터만을 포함하는 데이터가 그 예이다), 잡음이 있고(관심있는 속성 값이 없거나 집계 데이터만을 포함하는 데이터가 그 예이다), 일치성이 없다(관심있는 속성 값이 없거나 집계 데이터만을 포함하는 데이터가 그 예이다). 따라서 데이터 속의 내재된 의미나 패턴을 찾아 지식을 발견하기 위해서는 이러한 실제 생활 데이터를 보정할 필요가 있는데, 이렇게 데이터를 수정하는 과정을 데이터 전처리라고 한다.

결손 데이터의 결측치를 채워 넣고, 이상치를 식별하거나 제거하여 불일치를 해소하는 데이터 정제(data cleaning)와 여러 소스에서 온 데이터들을 하나의 통일된 형태로 융합하는 데이터 통합(data integration)이 대표적인 과정이다. 전술한 데이터 웨어하우스에서 데이터 마이닝을 적용하는 경우 데이터 웨어하우스 자체에 이러한 데이터 전처리 과정이 포함되어 있기 때문에 이 과정이 생략될 수도 있다[1].

3) 데이터 변환(data transformation, data coding)

분석의 효율성을 높이고 데이터의 복잡성을 감소시키기 위해 차원 축소와 같은 방법으로 고려해야 할 데이터 수를 줄이고, 정확한 모형을 도출하기 위해 원 데이터를 바탕으로 적절한 설명변수를 선택하고, 파생변수를 생성하는 등

효율적인 계산을 위한 자료 형태를 변환하는 과정을 말한다. 변수선택, 변수생성, 로그함수, 역함수, 지수함수, 백분위수 등이 이러한 과정에 필요한 작업으로 상위 개념 계층을 이용한 일반화, 정규화, 이산화(연속형 속성 값에 대해 속성의 범위를 여러 구간들로 나눔으로써 그 수량을 줄이는 것을 말한다. 예를 들어 나이라는 연속적인 개념을 10대, 20대 등의 구간으로 나누는 것이다) 등의 방법도 사용된다[4].

4) 데이터 마이닝

지식 발견 과정의 가장 핵심적인 과정이며, 좁은 의미의 데이터 마이닝을 말하는 것으로, 분석 목적에 따라 특정한 데이터 마이닝 기법을 선택하고, 적절한 알고리즘을 선정하여 내재된 의미나 패턴을 탐색하는 단계를 말한다. 구체적인 데이터 마이닝 기법에 대해서는 후술하도록 한다.

5) 해석 및 평가

위와 같은 과정을 통하여 도출된 결과를 해석하고, 해석된 결과를 실제 문제에 적용하여 이전 지식과의 유용성을 평가하는 단계이다. 결과를 보다 쉽게 이해하기 위해 적절한 시각화 기법의 적용이 효율적이며, 유의미한 결과가 도출될 때까지 이전 단계의 과정을 여러 번 반복하게 된다. 만약 적합한 모형이 없다면 변수 선택이 잘못 되었는지, 분석 알고리즘이 적절치 않았는지 검토하여 다시 위 과정을 반복해야 하는데, 그림 3, 4에서도 이러한 반복 과정을 명확히 명시하고 있다.

4. 데이터 마이닝의 기능

지금까지 데이터 마이닝의 개념과 그 과정에 대해 살펴보았다. 그렇다면 이러한 데이터 마이닝을 통해 우리는 어떤 의미와 패턴을 찾을 수 있는지 살펴보도록 한다.

데이터 마이닝은 데이터 속에서는 쉽게 찾을 수 없으나 유용한 패턴을 설명하는 데 유용하다. 일반적으로 이러한 데이터 마이닝 기능은 서술형(descriptive) 범주와 예측형(predictive) 범주로 나눌 수 있다. 서술형 범주는 확보된 데이터에 결과 값이 없는 경우(unsupervised data) 주어진 데이터를 설명하는 패턴을 찾아 데이터 사용자가 이해할 수 있도록 표현하는 것을 말하고, 예측형 범주는 확보된 데이터에 결과 값이 있는 경우(supervised data) 주어진 데이터를 통해 모델을 생성하여 새로운 문제에 적용 가능한 값을 예측하는 것을 말한다. 즉 서술형 마이닝 작업은 데이터 속에 있는 일반적인 특성을 설명하는 것이고, 예측형 마이닝 작업은 미래 의사 결정에 관한 예측을 위해 현재 데이터들로부터 추론을 수행하는 것을 말한다. 서술형 범주의 유형에는 개념/클래스화, 연관성 분석, 군집 분석, 이상치 분석, 전개 분석이 있고, 예측형 범주의 유형에는 분류 분석이 있다[1, 2].

1) 개념/클래스 기술(특성화와 차별화)

데이터는 클래스(데이터구조와 메소드들로 구성된 객체들을 모임을 클래스라고 한다)나 개념을 수반한다. 각각의 클래스와 개념들을 요약, 정리된 용어로 서술하는 것은 하위 개념을 파악하는 데 있어 매우 유용한다. 데이터 마이닝을 통하여 모호하고, 불확정된 어떤 클래스나 개념을 서술하는 것이 가능한데 이를 개념/클래스 기술(記述)이라고 한다. 이러한 서술은 다음 방법을 통하여 유도될 수 있다.

① 데이터 특성화(data characterization) : 목표 클래스의 데이터들을 일반적인 특성이나 특징을 요약하는 것을 말한다.

② 데이터 차별화(data discrimination) : 목표 클래스를 하나 또는 한 집합의 대조 클래스와 비교하는 것을 말한다.

2) 연관성 분석(association analysis)

연관성 분석은 주어진 데이터의 집합에서 빈번하게 발생하는 속성(attribute), 값(value)의 조건들을 나타내는 연관규칙(association rule)을 발견하는 것이다. 즉 데이터 내에 존재하는 친화도나 패턴을 찾아 연관성이 많은 것을 발견해내는 것이다.

3) 분류(classification)

분류는 일련의 범주들이 사전에 제시되어 있을 때, 특정한 데이터의 항목은 이러한 분류 체계 중 어디에 속하는 것을 밝히는 것을 말한다. 즉 데이터의 클래스나 개념을 설명하고, 구별하는 모델들의 집합을 찾는 과정이며, 그 모델을 이용하여 클래스의 레이블이 알려져 있지 않은 객체들의 클래스를 예측하는 데 그 목적이 있다. 기존의 데이터에서 추출한 훈련 데이터(training data)를 토대로 이를 분석하여 모델을 유도한 다음 새로운 레이블을 예측하는 것이다.

4) 군집화(clustering)

군집화는 쉽게 말하여 데이터 중 유사한 것들을 몇 개의 집단으로 그룹화하여 데이터 전체의 구조를 파악하는 것이다. 다만 분류에서는 클래스 레이블이 붙은 객체, 즉 데이터 내의 특성을 바탕으로 이를 분석하는 것인 반면 군집화는 이미 알려진 데이터 클래스의 레이블을 참조하지 않고 데이터 객체를 분석하여 데이터의 객체들을 서로 높은 유사성을 지니지만 다른 군집의 객체와는 상이하도록 형성한다는 점에서 서로 차이가 있다.

이러한 군집화를 통하여 데이터의 일반적인 형태나 모델을 따르지 않는 데이터 객체들을 선별할 수 있는데 이를 이상치 분석이라고 한다. 통상 이러한 이상치 분석은 지식발견 과정의 데이터 전처리나 변환 과정에서 잡음 데이터를 선별하여 이를 폐기시키기 위해 사용되는데, 카드 사용 고객의 데이터 중 도난 카드 사용자의 유형과 같이 몇몇 응용분야에서는 오히려 일반적이고 정형적 데이터 형태보다 더 가치가 있을 수 있다.

5) 전개 분석(evolution analysis)

데이터 전개 분석은 행위가 시간에 따라서 변화하는 객체들에 대한 규칙성이나 경향을 묘사하고 모델링하는 것을 말한다. 이 분석 방법은 시간과 관련된 데이터들의 특성화, 차별화, 연관성, 분류, 군집화를 포함하며, 연속 혹은 주기성 패턴(sequential pattern) 추출 등도 이에 포함된다.

5. 데이터 마이닝의 기법

지금까지 소개된 데이터 마이닝 기법들은 그 종류가 상당한 뿐 아니라 지금도 새로운 기법과 알고리즘들이 각 연구소를 통해 계속해서 연구되고 있다. 따라서 처음 이를 적용하려 하거나 실제 사례에 적용하고자 한다면 각자의 상황에 적합한 데이터 마이닝 기법을 선택한다는 것이 그리 쉬운 것은 아니다. 데이터 마이닝 작업 유형에 관계없이 가장 탁월한 성능을 제공하는 특정 기법이 존재하는 것도 아니고 유사 기법이라고 하더라도 분석 대상이 되는 데이터의 특성이나 도출하고자 하는 지식의 성격에 따라 각기 다른 결과를 낳기 때문이다. 또한 어느 한 기법들이 독자적으로 사용되는 것은 아니고 데이터 마이닝을 하는 데 있어 복합적으로 사용될 수 있다(data mining hybrid technique). 이에 본 논문에서는 가장 많이 사용되는 대표적인 기법들을 소개하고 이의 적용 예를 살펴보도록 한다.

1) 의사결정나무(decision tree)

의사결정나무(decision tree)는 의사결정규칙(decision rule)을 나무구조로 도표화하여 관심대상이 되는 데이터를 몇 개의 소집단으로 분류(classfication)하거나 예측(prediction)하는 기법이다. 즉 이러한 과정을 통해 도출된 결과를 바탕으로 새로운 데이터를 분류하고 해당 데이터의 적절한 값을 적용시키는 것이다[5].

이러한 의사결정나무는 나무 구조를 통하여 분류의 과정을 보여주므로 다른 기법들보다 과정을 쉽게 이해할 수 있다는 강점을 갖는다. 통계학적인 용어를 사용하지 않고도 각 데이터에 영향을 주는 변수와 변수들의 상호작용을 누구나 쉽게 이해할 수 있도록 설명 가능한 기법으로 대표적인 데이터 마이닝 기법이라 할 수 있다. 이러한 의사결정나무는 데이터 마이닝의 기능 중 분류(classification) 기능을 하고자 할 때 많이 이용되며, 경영 마케팅 분야에서 고객유치, 수요 및 판매 예측과 마케팅 전략, 이탈 고객 분석 등에 주로 사용된다.

구체적인 알고리즘으로는 1960년대 Morgan과 Songuist에 의해 고안된 AID(automatic interaction detection)를 시작으로 1970년대 Morgan과 Messenger에 의해 THAID(theta automatic interaction detection)로 발전하여 현재는 CHAID(Kass), CART(Breiman, Friedman, Olshen, Stone)과 C5.0(Quinlan)이 주로 사용되고 있다[6]. 구체적인 알고리즘에 대한 언급은 본 논문의 주제를 벗어나는 것이므로 생략하도록 한다.

의사결정나무 기법의 구체적인 단계는 다음과 같다. 우선 분류오류를 크게 할 위험이 크거나 부적절한 추론규칙을 가지고 있는 데이터를 제거한 후(가지치기), 분석 목적과 자료구조에 따라 적절한 분리기준과 정지규칙을 지정하여 의사결정나무를 얻는다. 마지막으로 이익도표나 위험도표 또는 검증용 자료와 교차타당성을 검토하여 의사결정나무로 도출한 결론을 검증한다.

의사결정나무의 장점은 중요한 입력변수를 찾기가 쉽고 모형이 간단하여 이해가 쉬우며, 결과 예측에 대한 적절한 설명이 가능하다는 것이다. 즉 어떤 데이터에서 추출된 지식이나 데이터 모델에 대한 해석력이 우수하다는 것이 가장 큰 장점이다. 뿐만 아니라 데이터 간의 상호작용과 비관계성을 자동적으로 찾아내는 알고리즘이며, 선형성, 정규성, 등분산성 가정이 필요 없고, 비모수적 모형에서도

활용 가능하다는 장점이 있다. 반면 의사결정나무는 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근방에서는 예측 오류가 클 가능성이 있으며 선형모형에서 주된 효과를 찾아내기는 곤란하며 훈련용 데이터에 의존하기 때문에 결과의 예측에 있어 불안정할 가능성이 있다. 따라서 검증용 자료에 의한 교차타당성 평가나 전술한 가지치기에 의한 안정성 있는 의사결정나무 기법을 활용하는 것이 바람직하다[2].

2) 신경망 모형(neural network)

신경망 모형은 인간이 경험으로부터 학습해 가는 두뇌의 신경세포를 모방한 개념으로 마디(node)와 고리(link), 연결가중치(weight), 입력부위(input node) 등을 바탕으로 은닉마디라고 하는 독특한 구성요소를 포함하는 망구조를 형성하여 과거에 수집된 데이터로부터 반복적인 학습과정을 거쳐 데이터에 내재되어 있는 의미나 패턴을 찾아내고 이를 일반화함으로써 향후 예측을 가능하게 하는 기법으로 데이터 마이닝의 분류, 군집화, 연관성 분석 발견과 같은 작업에 널리 이용된다[7].

신경망 모델의 시초는 1943년 워렌 맥컬럭(Warren Mcculloch)과 월터 피츠(Walter Pitss)로부터 시작되었다. 이들의 모델은 네트워크 내의 단순한 요소들의 연결을 통하여 무한한 컴퓨팅 능력을 가진다는 점을 제시하였고, 1949년 캐나다의 심리학자 도널드 헵(Donald Hebb)이 두 뉴런 사이의 연결강도를 조정할 수 있는 학습규칙을 발표하고, 1957년 프랭크 로젠블럿(Frank Rosenblatt)이 퍼셉트론(perceptron)이라는 최초의 신경망 모델을 발표하여 당시 상당히 고무적인 반응을 일으켰다. 그러나 1960년대 민스키(Minsky)와 페퍼트(Papert)가 이 퍼셉트론이 매우 간단한 규칙만을 학습할 수 있을 뿐 상당한 학습 시스템을 구축할 수 없다는 것을 수학적으로 증명하여 이후 20여 년간 침체의 길을 걷게 된다.[3] 그 후 퍼셉트론의 단점을 극복하여 네트워크 구성 시 한 개의 계층을 추가 구성한 다층 퍼셉트론 개념이 등장하게 되면서 신경망 연구는 새로운 계기를 마련하게 되었다. 즉, 은닉층(hidden layer)과 역전파(back propagation) 학습 알고리즘을 이용함으로써 선형 분리 문제뿐 아니라 여러 가지 문제점들을 해결할 수 있게 된 것이다[1].

신경망 모델의 가장 큰 장점은 다양한 문제에 적용 가능하다는 점과 일반화 성능이 의사결정나무 모델에 비교하여 정확하고 우수하다는 점이다. 따라서 신경망 기법은 경영 마케팅 뿐 아니라 문자인식, 음성인식, 주가 예측과 같은 금융 문제 등에 광범위하게 활용되고 있다. 또한 데이터의 잡음에 대해 비교적 견고하고 현재 다양한 소프트웨어 패키지가 개발되어 활용도 간편하다는 장점이 있다. 반면에 신경망 모델은 분류나 예측 결과만을 제공할 뿐 결과 도출에 대한 이유를 설명하기에는 부족하다는 단점이 있다.

3) 규칙 귀납(rule induction)

규칙 귀납은 통계적 중요도에 근거하여 데이터로부터 유용한 규칙을 추출하고, 각 레코드 셋에 대해 항목의 친화도나 패턴을 찾아내는 기법을 말한다. 쉽게 설명하자면 데이터간의 연관성 정도를 측정하여 관련성이 높은 것들을 밝혀주는 것으로 데이터 안에 존재하는 항목간의 종속관계를 찾아내는 작업이다. 통상 마케팅에서 고객이 교차구입하는 품목을 알아본다는 의미로 장바구니 분석(market basket analysis)이라고 불리고 있으며, 이러한 장바구니 분석은 두 품목의 동시 구매가 얼마나 자주 일어나는가를 나타내는 지지도(support = Pr(X and Y) = (품목 X와 Y를 동시에 포함하는 거래 수) / (전체 거래수)로 표현된다), 품목 하나가 구매되었을 때 추가로 다른 품목이 구매될 확률을 나타내는 신뢰도(confidence = Pr(X and Y) / Pr(X) = (품목 X와 Y를 동시에 포함하는 거래 수) / (품목 X를 포함하는 거래수)로 표현된다) 등에 의해 개체간의 유용한 관계를 이끌어낸다. 예를 들면 미국의 대형 마트 할인점에서 이러한 장바구니 분석을 실시한 결과 일회용 아기 기저귀와 맥주가 함께 구매된다는 흥미로운 결과를 도출하여 매장 진열에 이용하고 있다(이는 아기 기저귀를 구입하는 미국 남성의 전형적 소비 패턴에 기인한다고 한다. 즉 미국 남성들은 주로 아내의 심부름으로 퇴근길에 아기 기저귀를 구입하게 되는데, 이 때 자신이 저녁 TV를 보며 마실 맥주를 같이 구매한다는 것이다). 그러나 이러한 규칙이 모두 흥미로운 것은 아니다. 예를 들어 예전에 동일한 제조사의 전자제품을 주로 구매했던 고객은 신제품 구매에서도 동일한 회사의 제품을 구매한다와 같이 일반적으로 알아낼 수 있는 규칙은 의미가 없으며, 새로 문을 연 건축 자재점에서 변기덮개가 많이 팔린다와 같이 타당한 근거 없는 연관성은 아무런 의미가 없는 것이기 때문이다. 규칙 귀납은 데이터 마이닝의 연관성 분석이나 군집화에 주로 사용되며, 매장진열이나 첨부 우편, 사기 적발 시스템 등에서 다양하게 활용되고 있다. 주요 알고리즘으로는 apriori, PRIM(Patient Rule Induction Method, PRIM을 별도의 개별

기법으로 소개하고 있는 문헌도 있으나 PRIM은 다차원의 자료를 공간상의 상자 형태로 이해하고 그 상자를 제거해 가며 남은 자료 값에 대한 목적 함수 값을 계산하는 기법으로 규칙에 의한 군집화와 목적함수 값의 최적화를 나타내는 알고리즘을 채택하고 있으므로 귀납적 규칙을 따르고 있다고 판단하여 규칙 귀납(rule induction)의 기법으로 분류하였다) 등이 있다[6].

이러한 규칙 귀납은 관련이 높은 것들을 밝힘으로써 종국에는 조건반응(if-then, 위의 기저귀와 맥주의 경우 기저귀를 구매하면 맥주를 구매할 가능성이 크다와 같이 표현될 수 있다)으로 표현할 수 있게 되므로 데이터 마이닝의 기능 중 연관성 분석의 결과를 이해하는 데 아주 유용하다. 또한 분석방향이나 목적변수가 없는 경우에도 활용 가능하며, 거래내용에 대한 데이터를 변환 없이 그대로 적용할 수 있는 간단한 자료구조를 갖는 분석방법이다. 반면 품목수가 증가하게 되면 분석에 필요한 계산은 기하급수적으로 늘어난다. 너무 세분화된 품목을 가지고 규칙을 찾으려 할 경우 의미 없는 분석이 될 수 있으며, 빈도수가 적은 데이터의 경우 당연히 연관된 수가 적을 것이므로 규칙 발견 시 제외되기 쉽다는 단점이 있다.

4) K-평균 군집화(K-means clustering)

데이터 마이닝의 군집화 기능을 이용하고자 할 때 대표적으로 많이 사용하는 알고리즘으로 각 개체를 가장 가까운 중심점에 할당하는 방법으로 입력 값 K를 취하고, 군집 내 유사성은 높고 군집간의 유사성은 낮게 되도록 각 개체들을 K 군집으로 분해하는 것을 말한다. 군집 유사성은 군집에서 군집의 무게중심으로 볼 수 있는 개체들의 평균값으로 측정한다. 즉 입력된 K의 초기 중심점을 선택한 다음, 각 개체를 가장 가까운 중심점을 갖는 군집으로 할당한 후 새로운 군집의 중심점을 계산하고, 이러한 작업을 할당에 변화가 없을 때까지 반복 수행하여 최종적으로 K개의 군집을 형성하게 되는 것이다[1].

이러한 K-mean 군집기법은 대용량에 대한 탐색적인 기법으로 사전 정보 없이 의미 있는 자료구조를 얻을 수 있고 거의 모든 형태의 데이터에 적용 가능하며 변수들에 대한 역할정의가 필요 없으므로 적용이 쉽다. 따라서 금융기관에서 고객의 평균 예탁금, 월평균약정, 회전율, 거래건수 등을 바탕으로 고객세분화를 실시한 후 이탈과 복귀를 반복하는 고객군 선별하고 이들의 특성을 파악하는 데 자주 활용되고 있다. 그러나 조건을 만족시키는 비유사성 거리의 척도와 일반적으로 많이 사용하는 가중치 결정이 난해하고, 군집수 k를 적절히 제시하지 않으면 유용한 결과를 얻을 수 없으며, 결과해석이 애매하다는 단점이 있다.

5) 시각화(visualization)

시각화 작업은 독단적 기법으로 활용되기보다 KDD 마지막 단계에서 예측 결과나 각 기법 적용과정의 이해를 높이는 데 많이 사용되므로 통상 데이터 마이닝의 기법에 시각화 기법을 생략하는 경우도 많으나 후술할 데이터 마이닝 수사 적용에 있어 범죄 조직간 관계 파악에 유용하게 활용되므로 본 논문에서는 시각화기법을 따로 들어 설명하도록 한다.

시각화는 발견된 지식을 이해 가능하게 하고 해석 가능하게 만드는 데 있어서 중요한 역할을 한다. 인간의 시각-뇌 시스템은 지금까지 알려진 가장 훌륭한 형태 인식 도구이기 때문이다. 따라서 데이터 집합에서 패턴을 발견하는 데 매우 유용하다고 할 수 있다. 이러한 시각화 기법은 단순한 산포도나 막대그래프로부터 3차원 무비에 이르기까지 매우 다양하며 특히 다차원의 대량 데이터일 경우 이 시각화로 이해하지 못한다면 인간의 이해력의 한계를 넘는 경우가 많기 때문에 이러한 다차원 데이터를 시각화하려는 고난위도의 기법들이 계속 연구되고 있다.

이러한 시각화 기법은 다차원 데이터의 개념 파악을 하는 데 있어 개별 기법으로서 유용하게 사용되기도 하지만 데이터 집합의 질적인 면과 어디에서 패턴이 발견될 것인가에 대한 개략적인 판단을 하는 데도 매우 유용하므로 위에서 밝힌 지식발견(KDD) 단계의 초기 단계에서도 자주 사용되며, 각 데이터 마이닝 기법을 통하여 추출한 지식이나 패턴을 표현하는 데도 많이 사용되므로 지식발견(KDD) 마지막 단계에서도 자주 사용된다[8].

대표적인 기법으로는 두 항목의 정보가 동일 공간에 출력되는 분산 다이어그램(scatter diagram)과 데이터 집합에서 의미있는 투영(projection)을 탐색하는 투영 추적(projection pursuit ; 한 관측치 당 여러 개의 변수로 구성된 다차원의 형태를 가지는 데이터를 2차원 내지 3차원의 저차원 공간에 투영하여 변수들 간의 관계를 사람의 눈으로 쉽게 관찰할 수 있도록 하는 것으로 Friedman과 Tukey에 의해 소개되었다. 그러나 이러한 투영 추적은 시각화에서 뿐만 아니라 확률 밀도 함수, 추정, 분류, 회귀분석 등의 다양한 통계기법들과 접목되어 유용한 알고리즘을 만들어 내고 있다) 등이 있다[3].

Ⅳ. 외국 데이터 마이닝 수사 적용 연구와 사례 적용

1. 외국 현황

미국은 9.11. 테러 이후 테러 및 범죄수사에 관하여 시민의 자유권을 광범위하게 제약할 수 있는 애국법(Anti-terrorism legislation, 일명 Patriot Act)을 제정하고 반테러 활동을 강화하고 있다. 이러한 경향과 발맞추어 9.11. 테러 이후 FBI 국장 로버트 뮬러(Robert S. Mueller)는 국회 연설에서 “그동안 우리는 테러 발생 이후 사태 진압이나 수사에만 초점을 맞추었고, 이를 예방하는 것에는 무관심하였다. 하지만 지금부터는 테러가 일어나기 이전에 관련 정보를 바탕으로 사전에 이를 예측하고 차단하는 활동을 강화하겠다.”고 밝혔다.

2002년 6월 3일자 “Federal Computer Week”에서는 “Investigative Data Mining Part of Broad Initiative to Fight Terrorism”이라는 제목으로 FBI가 테러에 대비하기 위한 새로운 기술로서 데이터 웨어하우징(data warehousing)을 도입하고, 테러분자들의 활동에 관한 광범위한 자료를 다루기 위해 데이터 마이닝 기술과 분석 소프트웨어를 사용하기로 하였다는 기사를 발표하였다.

뿐만 아니라 연방 법집행 관련 기관 및 각 주 경찰은 대학 연구소와 연계하여 범죄와 범죄자에 관한 방대한 양의 자료를 처리하고, 이들 사이의 유용한 지식과 패턴을 추출하기 위해 많은 연구를 하고 있는데 그 대표적인 예인 COPLINK를 보면 다음과 같다[9].

미국 애리조나 주립대학 정보관리학부 인공지능 연구실(Artificial Intelligence Lab, Management Information Systems Department, University of Arizona)의 쉰천 첸(Hsinchun Chen)을 비롯한 연구진은 투산경찰국(Tucson Police Department, 미국의 경우 한국 경찰 시스템과는 상이하여 지방자치체내 한 부서로 조직되어 있는 경우가 많다. 투산의 경우도 이에 해당하기 때문에 경찰서라는 개념을 사용하지 않고 경찰국으로 표현하였다)과 공동하여 범죄자와 범죄 자료에 관한 데이터 웨어하우스를 구축하고 데이터 마이닝 기법을 적용한 COPLINK 시스템을 개발하여 활용하고 있다. 표 1은 이러한 COPLINK 시스템의 데이터에 대한 개요를 보여준다.

표 1. COPLINK 시스템의 개요.

Size of Database	1.5M criminal records (528 Mbytes)
Size of Rusulting Concept Space	1.24 M terms (478 Mbytes)
Number of Names	644,143 terms
Number of Addresses	210,003 terms
Number of Vehicles	361,126 terms
Number of Organizations	27,158 terms
Number of Weapons	96 termss
Number of Crime Types	719 terms
Processing Machine	DEC Alpha Server 4100
Processing Time	24 minutes

이 시스템에서 주목할만한 점은 미국은 우리나라의 주민등록번호나 지문체계와 같이 전 국민을 대상으로 한 데이터가 구축되지 않아 인적자료, 주거자료, 차량자료, 사건자료를 바탕으로 용의자를 추적하는 개인 식별에 초점을 맞추고 있다는 점이다. 특히 발음은 같더라도 이름의 철자가 상이한 영어 특성으로 텍스트 데이터 마이닝(text data mining)을 활용한 것으로 파악된다. 그림 1은 이러한 체계의 개념을 도식화하고 있다. 그림 2에서 그림 5는 GUI 환경에서 구축된 COPLINK 활용과정을 보여준다.

이러한 용의자 특정이나 개인식별 외에도 COPLINK는 위에서 밝힌 시각화 개념을 도입하여 조직범죄의 체계를 분석하는 데도 활용하고 있다. 그림 6은 60여 명의 마약 관련 범죄자들에 대한 체계를 밝히기 위해 범죄자들을 시각화한 다음(a), 데이터 마이닝의 군집화 기법을 이용하여 그 중심축을 파악하고(b), 각 범죄자의 조직 계급에 따라 분류하여 재구성한 후 57명의 마약 조직을 밝혀(c, d), 마지막으로 조직 범죄간의 외부 연결 고리를 파악한 것(e, f)을 단계별로 보여준다.

2. 사례 적용(보험 사기 범죄의 시각화)

지금부터 위에서 살펴본 데이터 마이닝 기법들을 이용하여 실제 사례에 적용하여 보도록 한다. 본 건에 이용된 사례는 실제 사기 건으로 접수된 것을 바탕으로 인적사항과 특정 회사의 명칭은 생략하고, 날짜는 재구성한 것임을 밝혀 둔다.

- 2000. 09. 27. 교통사고
- 2000. 10. 5. 사고자 초진, 보험회사 사고 접수
  (피해자 차량에서 미끄러진 사고라고 진술)

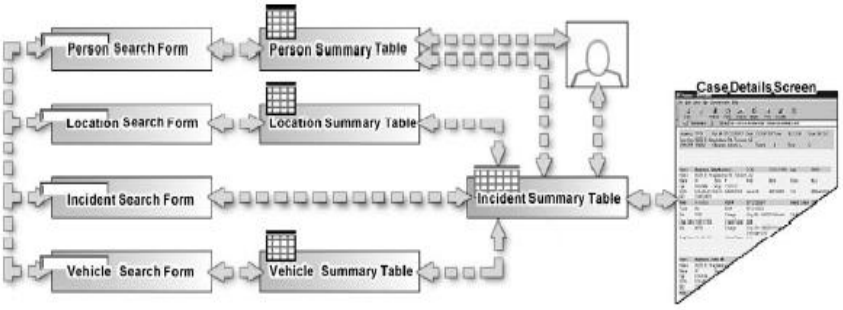

그림 1. COPLINK 적용 과정

그림 2. COPLINK 시스템의 인적 사항 검색

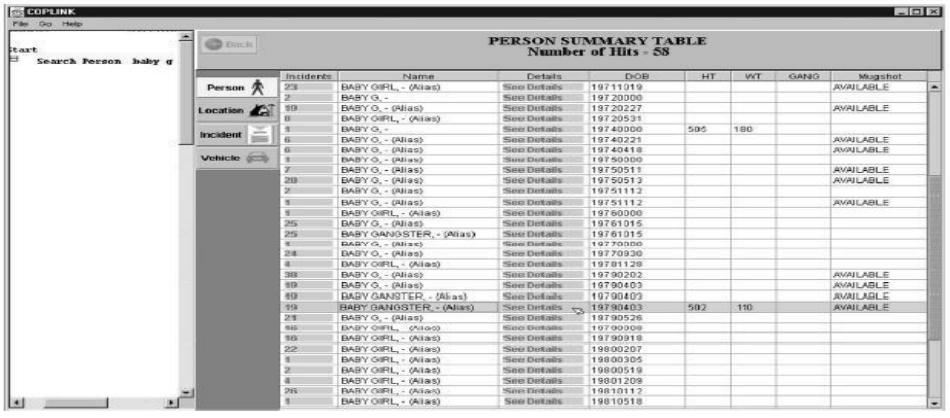

그림 3. COPLINK 시스템의 인적 검색 결과

그림 4. 인적 사항 내역

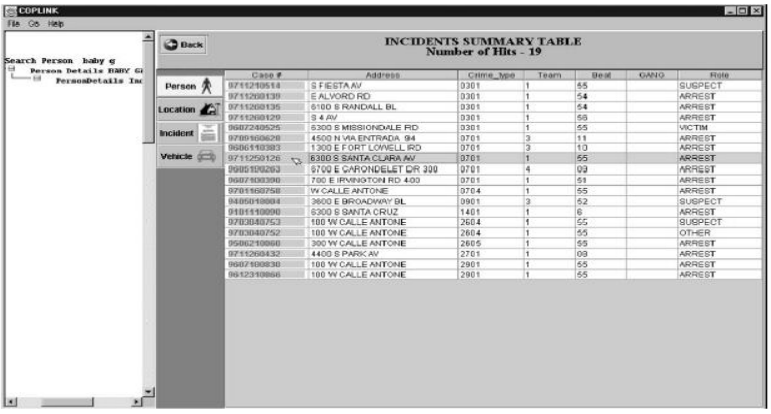

그림 5. 사건 기록 검색

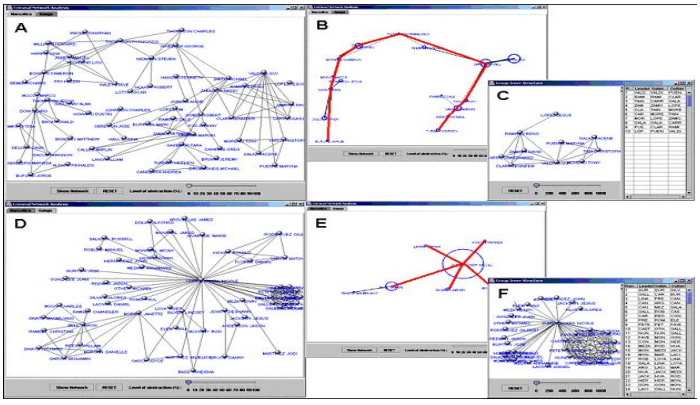

그림 6. 조직 범죄의 체계의 패턴 인식 과정

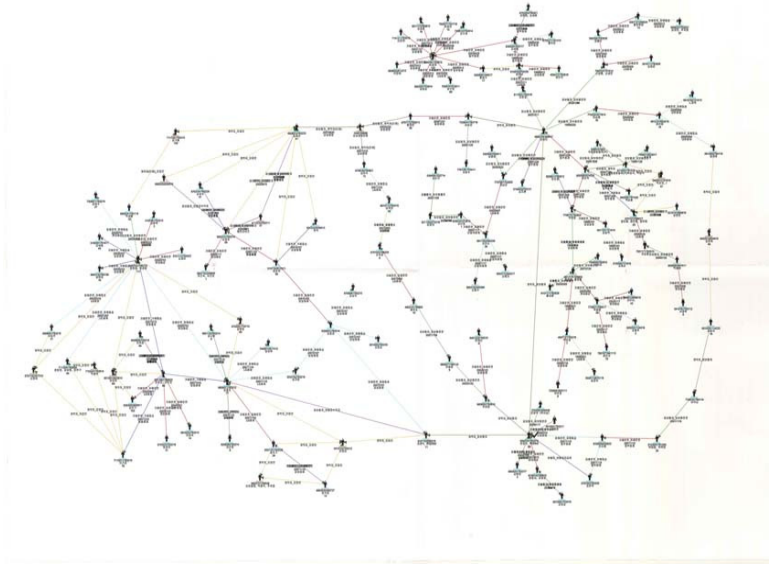

그림 7. 용의자 김○○를 중심으로 한 관련자 시각화

표 2. 용의자 김○○ 보험금 지급내역,

연번	사고일자	사고원인	역할	피해자수
1	99. 06. 15.	차선변경 추돌	피해운전	1
2	99. 11. 19.	갓길 진입 추돌	피해운전	7
3	00. 9. 22.	하차 중 단독사고	단독사고	1
4	05. 02. 23.	도로상 암석 충격	단독사고	1

표 3. 1999. 11. 19. 사고 인적 피해 및 보상내역.

(단위 : 일, 만원)

연번	피해자	부상급항	역할	피해자수
1	신○○	08-16	4	78
2	최○○	09-11	3	59
3	김○○ (용의자)	08-16	45	269
4	오○○	09-11	3	76
5	이○○	08-16	18	175
6	이○○	08-16	7	24
7	박○○	08-16	7	24

- 2000. 10. 7. A 보험회사 지급거절 (사유 : 단독 안전사고)
- 2001. 3. 12. 허리디스크 수술
- 2001. 6. 24. 보험금(진료비, 장해보험금 포함 3,200만원 지급)
- 2002. 1. 29. 보험사기 접수

사건 용의자 김○○는 2000. 9. 27. 교통사고 후 일주일이 지난 10. 5. 사고자 초진을 받고 그 후 아무런 재진이 없었던 점, 2000. 10. 7. 보험 지급 거절 후 5개월 반 정도가 지난 2001. 3월초 허리디스크 수술을 받고 2001. 6. 24. 보험금 지급된 점, 교통사고 이전 디스크 병력이 있는 점, 교통사고의 내용이 인적이 드문 장소에서 단독사고인 점 등으로 사기 혐의가 있다고 판단하여 금융감독원 보험조사실에 본 용의자의 과거 보험금 지급 내역을 의뢰하였다(표 2). 위 자료를 바탕으로 피해자수가 7명인 사건에 대해 다른 피해자의 인적피해 및 보상내역에 대한 자료를 다시 의뢰하였다(표 3).

위 표에서 볼 때 1999. 11. 19. 사고의 피해자 7명 중 용의자 김○○는 타 피해자와 비슷한 부상정도를 보이면서도 유독 장기입원하여 많은 보험금을 수령한 것을 파악하여 유력한 용의자로 판단, 위 용의자 김○○의 4건 사고와 관련된 모든 사람에 대한 관계를 데이터 마이닝 시각화 기법을 이용하여 제시하였다(그림 7).

용의자 김○○를 중심으로 보험모집인, 가해자, 피해자, 보험금 수령인, 피계약자, 수익자, 보험회사, 가해 및 피해 동승자 등의 관련자 117명에 대한 시각화를 통해 위와 같은 결과를 도출하였다. 전체적으로 하나의 닫힌 고리를 형성하면서 큰 내부 고리는 5개의 외부 작은 고리와 연결되어 있는 것으로 파악되나 내부 고리와 외부 고리의 관련성은 강하게 나타나지 않았다. 만약 한 사고의 가해자나 가해 동승자가 다른 사고의 피해자나 피해 동승자가 된다면 이러한 각 외부 고리의 개체는 내부 큰 고리와 더 많은 공유점을 나타낼 것이나 본 자료에서는 이러한 경향은 관찰되지 않았다.

그러나 사고 관련자 간에 전체의 닫힌 형태를 이루고 있는 점, 좌측 하단 부의 모험모집인간에 순환연결 이루고 있는 점 등을 볼 때 조직적 보험 사기의 가능성은 충분히 시사하며 이에 대한 수사가 이루어져야 할 것으로 판단된다.

Ⅴ. 데이터 마이닝의 수사 적용에 대한 검토

1. 효용성[10~12]

범죄 수사를 위해 많은 사람을 만나고, 많은 자료를 접하는 수사관들에게 방대한 양의 데이터를 분석하는 데 있어 데이터 마이닝은 분명 많은 가능성을 제시하고 있다. 특히 시각화를 바탕으로 조직 체계를 밝혀내는 기법은 외국의 예나 사례 분석에서 밝혔듯이 탁월한 효과가 있음을 확인하였다. 뿐만 아니라 과거 범죄와 범죄자들에 의한 데이터가 축적되어 효율적인 데이터 마이닝 기법이 적용된다면 향후 미제 범죄에 대한 중요한 단서가 제시될 수 있을 것으로 판단된다. 그 구체적인 활용방안은 다음과 같다.

1) 상습범죄에 대한 여죄 추적 및 분석

절도와 사기 등의 범죄는 습벽에 의한 상습성을 띄는 경우가 많으나 개별 사건에 있어 각각 범죄자를 특정하기에는 많은 어려움이 있고 미제로 남는 경우가 많다. 따라서 이러한 개별 미제 사건에 대해 데이터 마이닝의 군집화를 통해 큰 유형으로 분류한다면 향후 범인 검거시 효율적으로 활용될 수 있을 것이다.

2) 범죄자 프로파일링의 적용

무동기 연쇄살인이나 강간의 경우 범죄자와 피해자간의 대면이나 관련성이 없기 때문에 용의자를 추적하는 데 많은 어려움이 있다. 이에 수사기관은 이러한 무동기 범죄에 대해 과거 자료를 바탕으로 범죄자에 대한 프로파일링을 하고 있는데, 이 프로파일링에 데이터 마이닝의 연관성 분석과 군집화를 적용하면 많은 양의 데이터를 처리하는 데 효율적일 것으로 예상되며, 유사 범죄 발생시 이상치 분석을 통해 무동기 연쇄 살인이 맞는지, 맞다면 어느 범죄 유형과 일치하는지 파악하는 데 활용 가능하다.

3) 조직범죄의 체계 분석

마약판매나 자금세탁 등의 범죄는 한 개인이 범행하기에는 많은 어려움이 있어 조직적 활동을 필요로 한다. 수사기관이 사전에 이러한 조직적 범죄의 전체적인 윤곽을 파악하는 데는 많은 제약이 따르게 되는데 위에서 살펴본 바와 같이 통화 내역, 자금 거래 내역 등에 대한 데이터를 바탕으로 시각화와 군집화 기법을 적용한다면 내부 조직망이나 외부 조직과의 연계를 파악하는 데 유용할 것이다.

4) 패턴 인식을 이용한 자료 대비 시스템의 구축

데이터 마이닝의 신경망 기법은 해석력은 떨어지나 예측의 정확성과 다양한 문제에 대한 적용가능성, 잡음에 대한 견고성을 띄고 있다. 따라서 이러한 신경망 모형을 바탕으로 인지과학의 개념을 도입한다면 과거 전과자의 얼굴 화상 자료를 분석하여 군집화 하여 놓은 데이터베이스를 바탕으로 CCTV 등에 촬영된 용의자의 얼굴과 대비하는 시스템의 구축이 가능할 것이다.

2. 한계점 및 개선 방안

1) 데이터베이스 단계의 준비 미흡

수사기관은 사건을 수사하면서 많은 정보를 수집하게 된다. 현재 이러한 정보는 각 수사기관별로 데이터베이스화가 되어 있다. 경찰의 경우 범죄정보시스템에 의해 절도, 사기 등의 상습범죄, 구속피의자에 대한 수법자료, 용의자 불상 범죄에 대한 피해통보자료, 과거 수사 자료 등이 데이터베이스화되어 있다. 그러나 이러한 데이터베이스는 개별 항목으로 정리되어 있어 데이터 마이닝이 적용 가능한 환경을 지원하지 못한다.

따라서 위에서 밝힌 바와 같이 주제별로 통합된 데이터웨어하우스 형태로 데이터 관리 시스템을 변모시켜 각 파트별로 산재하여 있는 데이터를 활용 가능한 형태로 전환하는 정보화 작업이 먼저 요구된다.

2) 데이터 수집의 비정확성

수사기관이 수집하는 데이터는 범죄자나 사건 관련자들로부터 수집된 정보이거나 과거 발생된 자료이다. 그런데 범죄자들은 자신의 유리한 입장을 견지하기 위해 거짓말을 하고, 허위 정보를 유출하기도 한다. 이는 피해자의 경우도 마찬가지이다. 자신의 피해 품목을 과상 계산한다든지, 순간적으로 목격한 피의자를 다른 형태로 기술한다든지 하는 경우가 많다. 이렇게 부정확한 상황에서 과거 발생된 사안을 현재의 입장에서 수집하기 때문에 범죄나 용의자 데이터에 대한 정보가 부정확한 경우가 많다는 자체적인 한계점을 지니고 있다.

뿐만 아니라 이렇게 수집된 정보를 데이터베이스화함에 있어 현재 훈련되지 않은 개별 사건 담당자가 데이터를 입력하고, 마찬가지로 훈련되지 않은 전산 담당자가 데이터를 처리하는 경우가 많다. 따라서 표준화된 데이터의 형태로 축적이 되지 않고 있으며, 결손 데이터를 처리하는 데 어려움이 있다. 또한 마른 체격, 큰 키와 같은 부정확한 표현이 많아 데이터 마이닝을 적용하는 데 많은 제약이 따른다고 할 수 있다. 이에 각 데이터 수집에 표준화된 기준을 마련하고, 훈련된 요원에 의한 데이터 입력으로 충실한 데이터를 확보하는 것이 우선되어야 한다.

3) 데이터 수집의 한계(정보 독점화 우려와 국민의 기본권과의 충돌)

데이터 마이닝은 많은 양의 데이터를 축적하여 적용하여야 하며, 이를 반복하여 학습해 나가면서 지식을 발견하는 과정이다. 그러나 각 수사기관들이 보유하고 있는 자료는 각기 한정되어 있으므로 이를 연계하여 활용하지 않는다면 데이터 마이닝의 적용은 불가능하다고 할 수 있다. 예를 들어 경찰은 과거 전과자나 용의자에 대한 인적정보와 사건정보를 보유하고 있고, 재정경제부 소속 독립기구인 금융정보분석원(FIU)는 혐의 금융 데이터를 확보하고, 국세청은 재산 보유 현황 데이터를 확보하고 있는 등 각 기관마다 보유하고 있는 데이터는 상이하다. 범죄를 수사

하기 위해서는 각 기관의 데이터가 모두 바탕이 되어야 하며, 이 모든 데이터가 축적이 되어야만 비로소 데이터 마이닝을 통한 예측도 가능해지는 것이다.

그러나 이러한 데이터 마이닝의 수사 적용이나 수사 효율성을 위해 위의 모든 데이터를 한 기관이 독점하게 된다면 정보집중화 현상으로 권력 남용이 발생될 우려가 있다. 뿐만 아니라 과거 전과자나 용의자라 하더라도 한 개인의 인권과 사생활은 보장되어야 한다. 한 개인이 활동하는 데 있어 국가가 간섭할 권한도 없을 뿐더러 자신의 활동이 타인에게 알려져야 할 의무도 없기 때문이다. 따라서 수사기관이 방대한 데이터를 축적하는 데 내재적인 한계가 있다.

수사기관이 범죄 수사를 위해 데이터 마이닝 기법을 적용하기 위해서는 각 데이터를 확보하고 있는 정부기관, 민간부분에 대해 해당 사건에 관련된 사항만 조회가 가능하도록 법적인 제한이 이루어져야 한다. 다만 이 경우 원 데이터의 확보가 가능하도록 법적 준거가 마련되어야 할 것이다. 특히 민간인 혹은 금융기관, 통신기관과 같은 기업 등에 자료 요구는 압수수색영장이나 통신사실확인자료, 통신자료와 같은 엄격한 요건을 갖추어 개인의 인권이 무리하게 침해되는 일을 막아야 할 것이다. 이를 활용하고자 하는 수사기관은 각 사안별로 수집된 데이터를 통합, 재구성하여 데이터 마이닝을 적용하는 방향으로 접근하여야 하며, 각기 상이한 데이터 유형의 호환성을 높이는 방안도 함께 강구하여야 할 것이다.

뿐만 아니라 현재 과거 범죄나 전과자, 범죄자 관련 데이터베이스에 대해 수사기관에서 어느 정도까지 정보 수집이 가능하고, 데이터 저장이 가능한지에 대한 명확한 법적 근거가 없다. 이러한 한계의 모호성은 수사기관의 자의적 판단에 의하여 시민의 권리를 침해하는 결과를 초래하게 되므로 범죄와 범죄자에 대한 데이터베이스화에 대해 그 활용과 한계를 명확히 한 근거 법률을 제정함으로써 이에 따라 적용하여야 한다.

4) 예측 오류의 가능성(false negative 오류 위험성)

데이터 마이닝은 데이터를 해석하는 기능도 하고 있지만 본질적으로 현상을 예측하는 데 그 의미가 있다. 그러나 예측에 있어서 FN(False Negative), 즉 실제로는 범죄자가 아닌데 범죄자로 예측하고, 규정짓는 경우 그 위험성은 상당히 크다고 할 수 있다. 현대 형벌 제도는 수사기관인 행정부로부터 독립된 사법부에 의해서만 죄가 인정되고, 그 죄를 인정하기 위해서는 증거재판주의, 자백보강의 법칙 등 엄격한 요건을 충족시켜야 한다. 그런데 수사기관에서 이러한 데이터 마이닝의 활용으로 무리한 수사를 하다 보면 무고한 개인의 인권을 침해하는 경우가 발생할 수 있기 때문이다.

이는 데이터 마이닝을 주로 활용하는 기업의 마케팅, 경영 분야와는 확연히 구분되는 점으로서 기업의 경우 공격적 경영이 가능하므로 데이터 마이닝을 적용하여 예측 결과를 실제 경영에 활용한다고 하여도 별 무리가 없다. 고객 분류에 대한 구체적 분석이 이루어지지 않은 경우 10%의 고객이 이탈되었으나 데이터 마이닝의 고객 분류와 이에 대한 대응으로 5%의 고객만 이탈되었다면 기업의 입장에서는 충분한 효과를 거둔 것이기 때문이다. 그러나 수사에 있어서 예측 오류로 인한 결과는 개인의 인권과 생활에 치명적인 결과를 가져오는 것이기 때문에 데이터 마이닝을 수사에 적용하고자 할 때는 조심스럽게 접근하여야 한다.

따라서 수사기관은 데이터 마이닝의 결과를 범죄와 범죄자에 대한 단서로서의 의미로만 파악하여 수사개시 목적으로만 사용하여야 하며 다만 조직적 범죄유형이나 자금 세탁의 경우처럼 사후 범죄 개요를 파악하는 데 이용하는 것이 바람직하다고 생각한다.

참고문헌

1. Jiawei Han, 박우창 외3 역, 데이터마이닝 개념 및 기법, 2004.
2. 조재희, 박성진, OLAP 테크놀로지, 2003.
3. Pieter Adriaans, Dolf Zantinge, 용환승 역, 데이터마이닝, 1998.
4. 여윤희, Data Mining: review and comparison with statistical methods, 서울대학교 석사 학위 논문, 1999.
5. 진휘철, Data Mining은 우리에게 어떤 이득을 주는가?, 삼성SDS IT Review 2000. 7월.
6. 오희경, 데이터 마이닝 분류 모델 비교 및 분석, 서울대학교 석사 학위 논문, 2001.
7. Olivia Parr Rud, 이영섭 역, 데이터마이닝 Cookbook, 2003.
8. Jesus Mena, Investigative Data Mining for Security and Criminal Detection, 2003.
9. Hsinchun Chen et al, COPLINK Connect : information and knowledge management for law enforcement, Decision Support Systems 34, 2002.

10. Jennifer Xu, Hsinchun Chen, Criminal Network Analysis and Visualization : A Data Mining Perspective, 4th publication in Communications of the ACM, 2002.
11. Klerks, P. The network paradigm applied to criminal organizations : theoretical nitpicking or a relevant doctrine for investigator? Recent developments in the Netherlands, Connections, 24, No. 3, 2001.
12. Krebs, Mapping networks of terrorist cells, Connections, 24, No. 3, 2001.

초 록

데이터 마이닝은 컴퓨터와 정보처리의 발전으로 각기 다른 차원에서 다량으로 수집되는 데이터 속에서 숨은 의미나 패턴을 발견하는 유용한 기법이다. 의사결정나무, 신경망 모형, 규칙 귀납, K-평균 군집화, 시각화 등의 데이터 마이닝 개별 기법들은 산재해 있는 데이터에서 연관성을 분석하고, 이를 분류함으로써 일반화된 개념을 정의하고, 새로운 지식을 추론함으로써 실제 생활에 적용 가능한 예측을 가능하게 한다. 따라서 현재 데이터 마이닝은 기업의 마케팅 분야, 금융기관의 고객 분석, 통신 회사의 고객 이탈 방지 등에서 유용하게 활용되고 있다.

우리가 접해야 하는 정보의 양이 늘어나는 것은 범죄 수사에 있어서도 마찬가지 현상이다. 범죄와 범죄자에 대한 데이터는 축적되어 가지만 정작 개별 사안에 있어서는 중요한 데이터가 접근조차 되지 않고 있으며, 많은 데이터 속에서 이것이 내포하고 있는 숨은 의미를 지나치게 되는 경우도 많다. 본 연구에서는 선행 연구와 사례 적용을 통해 데이터 마이닝의 범죄 수사 적용 가능성과 한계점을 살펴보고자 하였다.

미제 사건으로 남는 경우가 많은 절도나 사기 같은 습관적 상습 범죄의 경우 데이터 마이닝의 분류, 군집화 기능을 활용한다면 향후 여죄 추적에 효율적으로 활용될 수 있음을 파악할 수 있었고, 특히 다양한 문제에 적용 가능하고, 잡음에 대한 견고성이 있음에도 예측의 정확성을 지니고 있는 신경망 모형의 경우 패턴 인식을 통하여 범죄자 프로파일링이나 화상 자료 대비 시스템 구축에 충분히 활용될 것으로 생각한다. 특히 보험 사기 사례 적용에서 살펴본 바와 같이 마약, 테러와 같은 조직적 범죄수사나 자금세탁과 같은 금융 추적 수사의 경우 해당 자료의 방대함과 모호성으로 인해 수사를 하는 데 많은 어려움이 있지만 이러한 데이터 마이닝 가시화 기법을 적절히 활용한다면 전체적인 윤곽을 파악하는 데 매우 유용하며, 효율적인 수사가 가능함을 확인할 수 있었다.

그러나 데이터 마이닝은 예측 모델이므로 오류를 내재하고 있다는 점에서 수사 기관의 데이터 마이닝 접근은 조심스러워야 하며, 정보 독점화 현상과 개인 사생활 보호라는 측면에서 각 수사기관은 해당 법률에 정한 범위 내에서 해당 사건별로 데이터를 수집하고 이를 통합, 재구성하여 활용하는 측면으로 적용되어야 할 것이다. 또한 각 수사기관별로는 자신의 보유하고 있는 데이터에 대해 다차원 처리가 가능하도록 데이터베이스 시스템을 구축하여 데이터 마이닝이 적용 가능한 환경을 구축하도록 하여야 할 것이다. 아직은 논의의 초기 단계이므로 효과가 크게 부각되지는 않았지만 지금까지 제시한 문제에 대한 연구가 계속 이루어진다면 인권중심, 증거중심의 수사 개념을 바탕으로 적법절차에 의한 수사 활동을 요구받는 시대에 새로운 대안으로 자리 잡을 것이며, 수사의 과학화에 기여할 것으로 전망한다.