https://doi.org/10.5392/JKCA.2021.21.02.121

관세 정형 빅데이터를 활용한 우범공급망 거래패턴 선별

Transaction Pattern Discrimination of Malicious Supply Chain using Tariff-Structured Big Data

김성찬*, 송사광*, 조민희, 신수현*****

한국과학기술정보연구원 연구데이터공유센터/UST 데이터 및 HPC 과학과*
한국과학기술정보연구원 연구데이터공유센터**,
관세청 관세국경위험관리센터***

**Seongchan Kim(sckim@kisti.re.kr)*, Sa-Kwang Song(esmallj@kisti.re.kr)*,
Minhee Cho(mini@kisti.re.kr)**, Su-Hyun Shin(sshlisa@korea.kr)*****

요약

본 연구에서는 데이터마이닝(Data Mining) 기법 중 하나인 연관관계분석(Association Rule Mining)을 적용하여 위험화물 선별모델을 구축함으로써 관세위험을 최소화하고자 한다. 이를 위해 관세청 수입신고서 빅데이터를 활용하여 연관관계분석 알고리즘인 어프라이어리 알고리즘(Apriori Algorithm)을 적용하고 공급망 간의 위험정도를 계산한다. 대규모의 수입신고 데이터로부터 해외공급자와 수입업체 간의 세율관련(과세가격, 품목, 중수량 등), 원산지표시 위반 등에 관련한 적발결과 관한 규칙셋(Rule Set)과 이 규칙들의 신뢰도(Confidence)을 확보하여 우범공급망 간의 거래패턴을 예측할 수 있는 선별모델을 구축한다. 총 2년 6개월치의 수입신고 데이터를 활용하여 5-겹 교차검증(5-fold cross validation)을 수행한 결과 16.6%의 Precision과 33.8%의 Recall을 보였다. 이는 빈도기반 방법보다 Precision 기준 약 3.4배 Recall 기준 약 1.5배 높은 결과이다. 이로써 논문에서 제안하고 있는 방법이 관세위험을 줄일 수 있는 효과적인 방법임을 확인하였다.

■ 중심어 : | 연관관계분석 | 관세위험선별 | 우범공급망 | 수입신고 데이터 | 정형 빅데이터 |

Abstract

In this study, we try to minimize the tariff risk by constructing a hazardous cargo screening model by applying Association Rule Mining, one of the data mining techniques. For this, the risk level between supply chains is calculated using the Apriori Algorithm, which is an association analysis algorithm, using the big data of the import declaration form of the Korea Customs Service(KCS). We perform data preprocessing and association rule mining to generate a model to be used in screening the supply chain. In the preprocessing process, we extract the attributes required for rule generation from the import declaration data after the error removing process. Then, we generate the rules by using the extracted attributes as inputs to the Apriori algorithm. The generated association rule model is loaded in the KCS screening system. When the import declaration which should be checked is received, the screening system refers to the model and returns the confidence value based on the supply chain information on the import declaration data. The result will be used to determine whether to check the import case. The 5-fold cross-validation of 16.6% precision and 33.8% recall showed that import declaration data for 2 years and 6 months were divided into learning data and test data. This is a result that is about 3.4 times higher in precision and 1.5 times higher in recall than frequency-based methods. This confirms that the proposed method is an effective way to reduce tariff risks.

■ keyword : | Association Rule Analysis | Tariff Risk Discrimination | Malicious Supply Chains | Import Declaration Data | Structured Big Data |

* 본 연구는 한국과학기술정보연구원 주요사업(국가연구데이터 공유·확산체제 구축 K-21-L01-C04-S01)의 지원으로 수행되었습니다.

접수일자 : 2020년 12월 08일 심사완료일 : 2021년 01월 14일
수정일자 : 2021년 01월 14일 교신저자 : 송사광, e-mail : esmallj@kisti.re.kr

I. 서 론

1. 연구목적 및 필요성

급변하는 관세환경 변화에 따른 사회안전과 국민건강 위협에 대해 우범화물에 대한 정확한 위험선별로 관세국경 선제 위험 차단을 위해 고도화된 위험선별기술이 필요하다. 테러물품 마약 등 위해물품이 수입이후 유통될 경우에는 치명적이고 회복 불가능한 피해를 초래할 수 있으므로 사회안전 및 국민건강 위협에 선제적이고 과학적인 대응이 가능한 지능기반 분석환경이 강화되어야 한다. 특히 해외공급자, 수입업체 등 우범공급망 간 거래패턴 선별은 우범공급망 간 거래에 대한 위험요소를 발굴하고 올바른 수입신고 및 부정 거래 방지를 위해 필수적이다.

관세당국은 빅데이터와 인공지능 등 4차 산업혁명 기술이 발전함에 따라 변화에 맞추어 관세행정 시스템에 ICT(Information & Communication Technology) 기술을 적용하여 관세행정체계를 구성하기 위한 다양한 방안을 검토하고 있다[1]. 통상적으로 수입통관절차는 [그림 1]과 같이 화물도착(Port Entry), 수입신고(Import Declaration), 검사대상 화물선별(Cargo Selectivity), 물품검사(Goods Inspection), 사후조치-관세납부(Post-Duty Payment), 반출(Release of Goods) 등의 절차가 진행된다. 일반적으로 통관절차의 검사과정(Goods Inspection)에서 증가하는 수입신고 물품 대비 검사관의 부족으로 전량검사를 수행하는 것은 현실적으로 불가능하다. 따라서 검사 전 잠재적으로 위험성이 높은 검사대상 화물선별(C/S)과정을 두어 통계 및 부대정보 등을 통해 지정한 화물 및 무작위로 선택하고 일부 화물만 검사한다. 화물선별은 우범화물 자동선별 시스템을 활용하여 수출입 되는 물품 중에서 전산에 미리 등록된 기준에 따라 우범가능성이 높다고 예상되는 물품 및 화물을 선정하는 것으로 선정된 물품과 화물만 집중적으로 검사함으로써 검사의 효율을 높이는 검사 관리기법이다[2]. 하지만 화물선별 시 단순 통계나 정보를 이용하여 효과적인 선별이 이루어지지 않는 문제점이 발생하고 있으며[1], 이를 해결하기 위하여 빅데이터에 기반하여 데이터마이닝 및 인공지능, 블록체인 기술을 활용한 스마트 통관체제 구축으로 4차 산업혁명 시대의 관세행정 실현을 위한 움직임이 나타나고 있다.

본 연구에서는 데이터마이닝(Data Mining) 기술을 활용하여 위험화물 선별모델을 개발하고 운용시스템에 제공함으로써 관세위험을 최소화하고자 한다. 이를 위해 관세청 수입신고서데이터(Import Declaration Data)를 대상으로 연관관계규칙 마이닝(Association Rule Mining) 기법을 적용하여 수입신고서에서 위험성을 내포한 규칙(Rule)을 선별한다. 선별된 규칙셋(Rule set)과 위험도를 포함하는 모델을 구축한 후 해외공급자, 수입업체, 수입물품 등에 관련된 우범공급망 간 거래패턴을 분석하여 세율관련, 원산지표시 위반 등 고위험 거래에 대한 수입 건(신고서)을 예측한다. 예측된 위험 물품과 화물은 화물선별(C/S)단계에서 수입통관 자동선별 시스템에서 검사관에게 통보되고 검사관은 수작업 정밀 검사를 진행하게 된다.

2. 인공지능 기반 우범화물 검사대상 선별 플랫폼

관세청은 대내외적으로 증가하는 무역량에 따른 업무량 증가와 검사율 하락 등에 유연하게 대처할 수 있는 머신러닝, 딥러닝 등 인공지능 기법을 활용한 시스

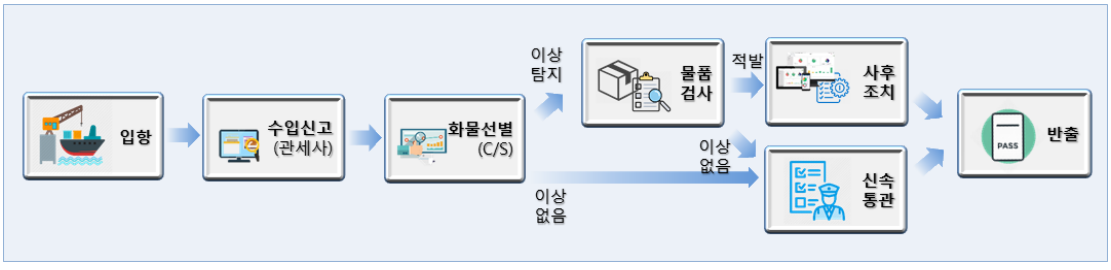

그림 1. 일반적인 수입통관 절차

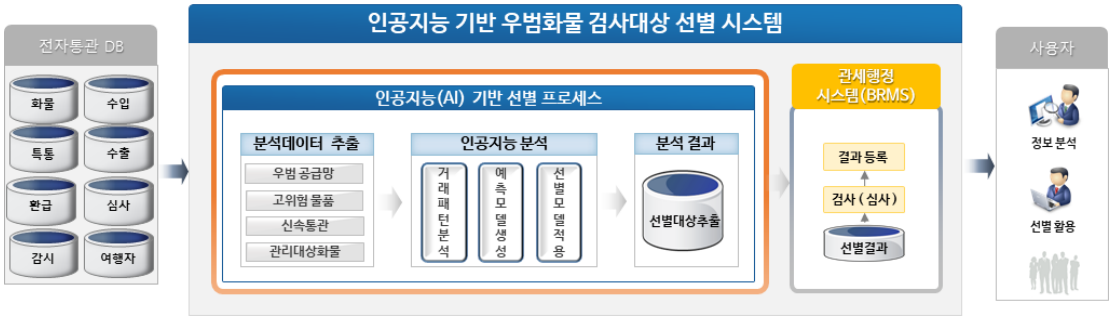

그림 2. 인공지능 기반 우범화물 검사대상 선별 시스템

템 구축을 추진하였다[3]. [그림 2]는 관세청에서 추진한 인공지능 기술 기반의 우범화물 검사대상 선별 플랫폼의 구체적인 모습을 보여주고 있다. 선별 플랫폼은 전자통관시스템의 과거에 축적된 수출입 통관자료, 무역 공급망 데이터를 수집하여 추출, 정제, 변환 등의 분석데이터 추출 및 전처리 가공을 거쳐 인공지능 예측모델 생성에 활용한다. 학습데이터와 인공지능 알고리즘 라이브러리를 활용해 예측 모델을 생성하고 모델평가를 포함한 최적화 과정을 거쳐 최종 모델을 구축한다. 구축된 선별모델은 AI 선별관리 시스템에 탑재되어 실제 운용되는 관세행정(전자통관)시스템에 신고된 수입신고건에 대해 선별을 진행하게 된다. 선별 플랫폼에서는 우범공급망 간 거래패턴 선별모델 뿐 아니라 고위험·저위험 물품에 대한 수입통관 선별모델, 관리대상화물 선별모델 등 다양한 모델을 탑재하여 운영한다. 추가적으로 자동 리모델링 시스템을 구축하여 선별모델을 운영함과 동시에 지속적으로 축적되는 데이터를 활용하여 모델을 지속적 보완함으로써 장애발생 최소화 및 최적 시스템 운영이 가능하도록 한다.

관세청은 인공지능 기술 기반 자가 학습을 통해 항상 최적의 상태 유지가 가능한 다양한 인공지능 기반 우범화물 검사대상 선별모델 운영을 통해 고위험 우범 수입화물의 적발률을 향상시키고, 저위험 화물을 빠르게 통관하도록 하여 정상화물에 대한 불필요한 검사 최소화한다. 이로써 신속통관이 가능하게 되어 및 검사비용 및 물류비용을 절감하여 수입업체의 무역행정 부담을 줄이는 효과가 있게 된다.

관세행정시스템- BRMS(Business Rule Management System)에 탑재되는 여러 선별모델 중 우범공급망 간 거래패턴 선별모델(Supply Chain Model)은 공급망 간 연관관계분석 후 규칙(Rule)을 생성하여 특정 적발과 관련된 공급망 간 연관관계를 선별한다. 학습데이터 구성은 수입거래자료 등 관세청이 보유한 데이터를 바탕으로 공급망 간 숨겨진 우범집단 도출이 가능하도록 공급망 학습데이터를 구성한다. 그 후 수입금지품, 마약류, 지재권 등의 적발과 관련된 공급망 별 상관관계 및 위험도를 분석하여 상관도에 따라 검사대상으로 자동 선별하는 모델이다.

II. 선행연구 및 이론적 배경

1. 선행연구

화물선별 효율화 방안이나 시스템에 관련한 연구들을 살펴보면, 최현철[4]은 일반 여행자들이 세관을 의식하지 않고 신속하게 통관할 수 있도록 우범여행자와 우범화물을 과학적인 기법으로 탐지하는 방법을 제안함과 동시에 통관제도를 개선 방안을 제안하였다. 영리목적으로 하지 않는 여행자휴대품의 면세범위를 확대를 기반으로 휴대품 검사관실의 업무 부담 개선 및 근무 체제 변경의 개선방안을 제시하였다. 권은주[5]는 관세행정 위험관리제도의 효율적 운영방안에 관한 연구에서 위험관리를 위해 빅데이터와 AI를 활용하는 방안에 대하여 논의하였다. 빅데이터와 AI를 활용하여 다양화 된 우범요소를 탐지할 수 있는 기능 및 모델을 개발하고 이를 관세청 통합위험관리시스템(IRM-PASS)에 적용해야 하는 필요성을 논의하였다. 정분도[6]는 전자통관시스템 활성화 방안에 관하여 연구하였으며, 선진

국 주요 국가의 전자통관시스템 추진 현황 대비 국내의 전자통관시스템의 개선점 도출 및 방안을 제시하였다. 주요 국가 및 국제기구에서의 한 창구에서 모든 관련 업무를 일괄 처리하는 싱글윈도우 구축 사례를 제시하고 한국 전자통과시스템에 싱글윈도우 도입을 제안하였다. 이명구[1]는 4차 산업혁명 기술을 활용해 관세행정에 적용하여 개선방안에 대한 연구를 수행하였다. 화물성격과 수출입업자의 과거 전력 등을 분석해 우범성을 탐지의 필요성을 제안하였고, 블록체인 기술을 활용하여 통관서류 통용이 가능하도록 하는 방안에 관하여 연구결과를 도출하였다. 송선욱[7]은 국제항공화물 운송기를 대상으로 하는 미국의 항공화물사전검색(ACAS) 프로그램을 연구하였다. ACAS는 항공기 적재 사전에 화물정보를 제출 받아 위험평가를 진행하도록 하는 것이다. ACAS는 국제적 협력 기반으로 항공화물 공급망 보안향상 측면에서 긍정적 역할을 하는 것으로 평가된다.

또한, 서동민[8]은 수입신고서 및 연관된 데이터가 저장된 관계형 데이터베이스에서 수출입 공급망 연계정보를 추출하여 네트워크 탐색기술로 무역거래 위험요소를 적발하고자 하였다. 또한 무역거래 위험요소를 가시화하여 전체 네트워크 및 키워드를 사용자에게 제공하고 유형 검색 기능을 제공하였다. 무역 거래 정보의 연계를 통한 연관 관계 중심의 위험 정보 분석 및 변화 추이 모니터링 기능을 제공함으로써 선별을 쉽게 할 수 있도록 하였다. 하지만 이 연구는 제안한 시스템에 대해 실제 데이터를 활용해 검증이 이루어지지 않아 시스템 효용성 측정이 불가능하다.

문헌조사 결과 전자통관시스템에 관련된 개선방안 제시에 대해서는 많은 연구가 이루어지고 있으나 실제 선별모델 구축 관련한 연구의 사례는 미미했다. 더욱이 구축된 선별모델이 실제 운용 시스템에서 실 수입 데이터로 성능을 검증한 사례는 이 연구가 처음으로 제시하고 있다.

2. 이론적 배경(연관규칙 마이닝)

연관규칙(Association rule) 학습은 대형 데이터베이스에서 변수 간의 흥미로운 관계를 발견하기 위한 규칙기반의 기계학습 방법이다[9]. Agrawal 등[10]에 의해 강력한 규칙의 개념을 바탕으로 슈퍼마켓 POS (point-of-sale) 시스템에서 기록한 대규모 거래 데이터에서 제품 간의 규칙성을 발견하는 연구가 시작되었다. 연관규칙 마이닝은 장바구니 분석(Market Basket Analysis)에 대한 예제 외에도 웹 사용 마이닝, 침입 탐지, 생물정보학을 비롯한 많은 분야에서 사용된다.

연관규칙 마이닝에서는 통계적 척도를 사용하여 데이터베이스에서 발견된 강력한 규칙을 식별하고 그 규칙을 연관규칙 $X \rightarrow Y$ 로 나타낸다[9]. 발견된 연관규칙은 다음 세가지의 척도로 성능을 평가한다. 연관규칙 $X \rightarrow Y$ 의 지지도(Support)는 전체 거래(transaction) 빈도(n) 가운데 아이템 X, Y가 동시 출현(freq(X,Y))한 비율을 나타내며 (1) 같이 표현한다. 신뢰도(Confidence)는 아이템 X가 출현한 빈도 가운데 아이템 Y가 함께 나타난 비율을 의미하며 (2)와 같이 표현된다. 향상도는 아이템 Y가 출현한 비율에 비해 아이템 X와 Y가 함께 나타난 비율을 의미하며 (3)과 같이 표현된다.

$$Support = P(XUY) = freq(X,Y)/n \quad (1)$$

$$Confidence = P(Y \mid X) = (P(Y \cup X))/(P(X)) = (Support)/(P(X)) \quad (2)$$

$$Lift = (P(Y \mid X))/(P(Y)) = (P(Y \cup X))/(P(X)P(Y)) = (Confidence)/(P(Y)) \quad (3)$$

Ⅲ. 우범공급망 거래패턴 선별모델 구축

본 장에서는 해외공급자, 수입업체 등 우범공급망 간 거래패턴을 탐지하는 선별모델을 기술한다. 우범공급망 선별모델은 과거 축적된 수입신고서 데이터를 활용해 우범공급망 정보를 바탕으로 수입신고건의 위반 여부를 예측하는 모델이다. 수입신고서의 공급망 정보로는 수입업체-해외거래처 뿐 아니라 거래품명, 신고품명, 수입종류코드, 원산지 등 기타 정보도 고려 대상으로 한다. 수입검사결과코드는 모델이 예측하고자 하는 정답(Ground-Truth)으로 활용한다.

[그림 3]은 우범공급망 선별모델의 생성과정 및 활용절차를 나타낸다. 우선 우범공급망 선별에 활용할 모델

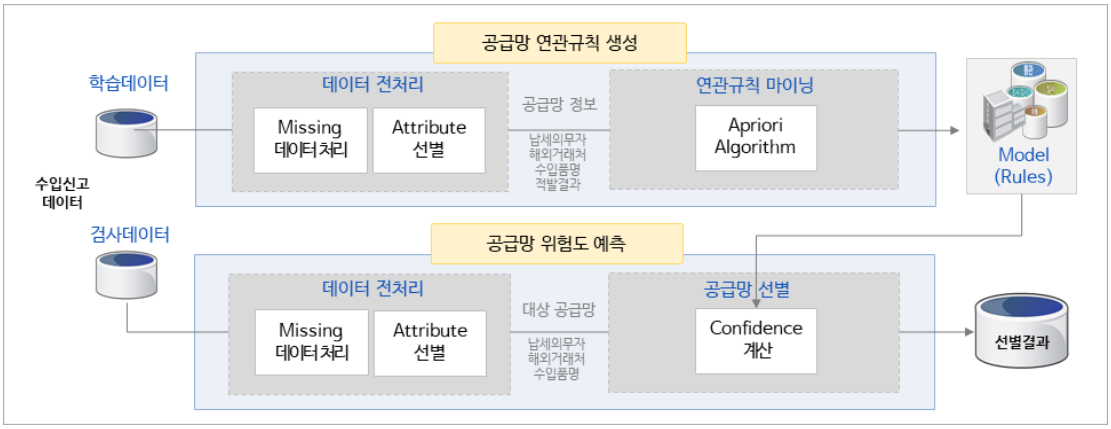

그림 3. 우범공급망 연관규칙 모델 생성 및 위험도 예측

(Rule Set)을 생성하기 위해 데이터 전처리(Preprocessing) 및 연관규칙 분석(Association Rule Mining)을 실시한다. 전처리 과정에서 오류처리를 거쳐 규칙(Rule)생성에 필요한 항목(Attribute)을 수입신고서 데이터의 추출한다. 추출한 항목들을 어프라이어리(Apriori) 알고리즘의 입력으로 하여 규칙들을 생성한다. 생성된 연관규칙 모델(Model)은 관세청 선별시스템에 탑재되어 검사여부를 판단해야 하는 수입신고서가 들어왔을 때 규칙과 비교할 항목들을 추출하여 모델을 참조하게 되고 공급망 기준으로 신뢰도 값을 반환하여 반환된 신뢰도 값이 검사여부의 기준점을 넘는지 비교하게 된다. 결과에 의해 해당 수입 건을 검사할지를 판별하게 된다.

1. 데이터 및 전처리

모델구축에는 2016년도부터 2018년도 상반기까지 2년 6개월간 수입 신고된 데이터 중 검사관들이 직접 검사를 실시하여 적발결과(수입검사결과코드)가 존재하는 644,583건의 데이터를 활용하였다. 정답 클래스로 여겨지는 수입검사결과코드[11]는 96개의 코드로 있으며, 대표적으로 '이상없음(no abnormality)', '원산지 허위표시', '상표권 침해', '동일세번내 저세율로 변경' 등의 적발코드가 있다. 전체 검사데이터 중 70.2%(452,243건)이 검사결과 '이상없음'의 수입 건이었고 나머지 192,340건(29.8%)이 39개의 적발코드가 부여된 수입 건이었다. 이 중 가격·운임·보험 등의 '누락 과표과세'(M5) 적발이 가장 많은 것으로 나타났으며, '기타 신고항목 오류정정'(C2)이 두번째로 많은 것으로 파악되었다.

수입신고서데이터(Import Declaration Data)는 수입신고서번호(Import Declaration Number), 신고일자(Date) 등 44개의 항목으로 구성되어 있다. 이 데이터에서 전처리를 거쳐 공급망 관점에서 연관관계분석을 위해 필요한 4개의 항목: 납세의무자(Taxpayers Business Registration Number), 수입검사결과적발코드(Inspection Result Code), 해외거래처부호

표 1. 전처리 후 샘플 데이터의 예

Import Declaration Number	Date	Taxpayers Business Registration Number	Inspection Result Code	Overseas Account Number	Trade Product Name
A124416A000001	20160519	A000015910	A01	A046067	OTHER
A124416A000002	20160321	A000106489	A65	A086034	CARTON
A124416A000003	20160322	A000051045	A01	A056476	SHIRT
A120916A000004	20161209	A000078955	A01	A126932	GARMENTS

(Overseas Account Number), 거래품명(Trade Product Name)만 추출하였다. 전처리로 결측치 및 이상치를 제거하는 데이터 오류처리를 거쳤다. [표 1]은 수입신고서 데이터 중 규칙생성에 추출한 항목들의 예제를 보여주고 있다.

2. 어프라이어리(Apriori) 알고리즘을 이용한 연관 관계분석

연관규칙 분석은 빈발 아이템 집합과 연관규칙을 찾는 작업으로 빈발 아이템 집합을 찾은 후 연관관계를 찾는다[9]. 본 연구에서는 후보집합을 찾는 어프라이어리(Apriori) 알고리즘을 이용하였다. 이 알고리즘은 [그림 4]와 같이 빈발집합 마이닝(Frequent Set Mining) 방법 중 하나로 어떤 아이템셋이 빈발하면 모든 서브셋(Subset) 또한 빈발하다는 전제로 후보를 줄인다. 따라서 빈발하지 않은 아이템셋의 서브셋들을 모두 가지치기(Prune) 함으로써 아이템집합이 기하급수적으로 늘어나는 것을 막을 수 있으며 수행시간을 줄일 수 있다. 어프라이어리 알고리즘은 최소지지도(minSupport)와 데이터 집합을 입력으로 받아 한가지 아이템이 포함된 모든 후보 아이템 집합의 목록에서 최소지지도를 만족하지 못하는 아이템 집합은 버리고 모든 집합이 버려질 때까지 되풀이한다. 데이터 집합을 살펴보는 함수에 대한 의사코드는 [그림 5]와 같다[12].

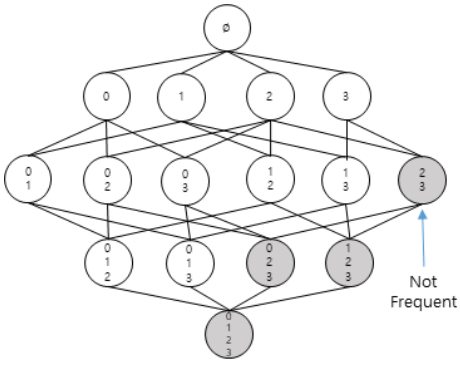

그림 4. Apriori 알고리즘 동작 예: 빈발하지 않는 하위 아이템셋도 역시 빈발하지 않으므로 고려하지 않음

```
#input: minSupport, dataset
For each transaction in tran the dataset:
For each candidate itemset, can:
        Check to see if can is a subset of tran
        If so increment the count of can
For each candidate itemset:
If the support meets the minimum, keep this item
Return list of frequent itemsets
```

그림 5. 후보 아이템셋를 생성하기 위한 의사코드

어프라이어리 알고리즘은 데이터 집합과 지지도가 제공되면 후보 아이템 집합으로 구성된 리스트를 생성한다. [그림 6]은 전체 어프라이어리 알고리즘의 의사결정 코드를 보여준다[12].

```
While the number of items in the set is greater than 0:
        Create a list of candidate itemsets of length k
        Scan the dataset to see if each itemset is frequent
        Keep frequent itemsets to create itemsets of length k+1
```

그림 6. 어프라이어리 알고리즘 의사코드

본 연구에서는 모델을 최소지지도(minSupoort)는 0.0001, 최소신뢰도는 0.1로 설정하고 모델(Rule Set)을 만들었다. 최소지지도를 0.0001로 설정하였으므로 만일 데이터셋의 transaction 개수가 500,000개라고 하면 50번 이상 출현한 아이템(혹은 Rule)만 연관규칙의 후보로 고려한다. 최소신뢰도 0.1을 설정하였으므로 생성된 규칙들 중 신뢰도가 0.1 미만인 규칙들은 버리게 된다.

연관규칙 모델을 구축 후 위험 선별에 사용할 때[그림 3 공급망 위험도 예측 참고]는 납세의무자(e.g., A000104861)와 해외거래업체(e.g., A070816)를 넣어 해당 납세의무자와 해외거래업체가 포함된 모든 규칙(rule)을 추출한다. 이 규칙 중 적발코드가 포함되지 않은 규칙은 모두 버리고 신뢰도 기준으로 내림차순 정렬하여 반환한다. 검사관들은 반환된 규칙 중 상위 N개의 결과를 참고하여 어떤 물품을 검사를 수행할지를 결정하게 된다.

예를 들어, 규칙 {(A000106489, A086034), (A02),

0.78} (A000106489, A086034 → A02)이 있다면 두 아이템 (A000106489, A086034)과 A02는 통계적으로 연관관계 있고 그 연관정도는 0.78이라는 의미한다. 즉, 납세의무자 A0001064890와 해외거래처 A086034는 A02 라는 적발코드에 대해 신뢰도 0.78로 우범성이 있다는 것을 의미한다. 실제 선별시스템에서는 반환되는 신뢰도를 기준으로 검사여부를 설정하여 검사를 진행하게 된다.

Ⅳ. 실험

1. 평가방법 및 결과

2016년도부터 2018년도 6월까지의 2년 반 기간의 전체 데이터 644,583건 데이터셋을 [표 2]와 같이 5겹 교차검증(5-fold)로 나누어 실험을 진행하였다. Fold-1의 경우 학습데이터 수는 523,273건이었으며 이중 약 30%정도가 결과코드가 '이상없음'이었다. 학습은 이를 제외한 나머지 데이터로 진행하였으며, 최소지지도(minSupport)는 0.0001로 하고 최소신뢰도(minConf)는 0.1로 하였을 때 학습 시간은 약 8.08시간이 소요되었다. 최소지지도와 최소신뢰도는 0.0001, 0.1로 설정했을 때 Fold-1에서 생성된 연관규칙셋(Rule Set) 수는 2972개였다.

구축한 모델의 성능 비교를 위해 빈도기반(Frequency-based) 추천방법과 비교하였다. 빈도기반 추천방법은 학습 데이터셋(학습셋)의 거래기록에서 하나의 납세의무자(수입업체)를 선택했을 때 가장 많이 거래한 해외거래처를 찾아 이 거래처와 가장 많이 위반한 적발코드를 추출하여 이를 코드를 추천하는 방법이다. 납세의무자-해외거래처 리스트에서 각 적발코드를 카운트하여 적발 코드 빈도 데이터베이스에 저장하여 놓고 가장 많은 빈도를 기록한 적발 코드를 반환하도록 구성하였다. 이후 테스트셋에서 검사하고자 하는 데이터에서 납세의무자-해외거래처를 추출한 후 이 적발코드가 정답과 맞는지를 평가하였다. 납세의무자-해외거래처 빈도기반 추천 데이터베이스에 없을 경우 데이터셋 거래기록에서 다음과 같은 순서대로 가장 많이 위반했던 적발코드를 반환하여 평가하였다.

1) 수입업체에서 제일 많이 위반했던 적발코드 반환
2) 해외거래처에서 제일 많이 위반했던 적발코드 반환
3) 거래품명으로 제일 많이 위반했던 적발코드 반환
4) 위에 해당되지 않을 경우 데이터셋 거래기록에서 가장 많이 위반한 적발코드 반환

구축한 모델의 평가 또한 모델에서 제시한 적발코드가 없을 시 위의 순서대로 적발코드를 산출하여 평가를

표 2. 데이터셋 및 실험결과

방법		분류	기간(년)	데이터 수	연관규칙 수 (minSup 0.0001/minConf 0.1)	Precision	Recall
Apriori	Fold 1	학습셋	2016-2017	523,273(81%)	2972	0.08	0.25
		평가셋	2018년 상반기	121,310(19%)			
	Fold 2	학습셋	2016, 2017 상반기, 2018	484,371(75%)	2600	0.2	0.39
		평가셋	2017 하반기	60,212(25%)			
	Fold 3	학습셋	2016, 2017 하반기, 2018	25,751(81%)	2616	0.19	0.38
		평가셋	2017 상반기	18,832(19%)			
	Fold 4	학습셋	2016 상반기, 2017, 2018	99,652(78%)	2626	0.2	0.41
		평가셋	2016 하반기	144,931(22%)			
	Fold 5	학습셋	2016 하반기, 2017, 2018	524,435(82%)	2846	0.16	0.26
		평가셋	2016 상반기	120,148(18%)			
	Avg				2732	0.166	0.338
빈도기반 (Baseline)	-				-	0.049	0.217

실시하였다. 빈도기반의 Precision은 0.049였으며, 5겹 교차검증 평가 결과는 [표 2]와 같다. 5-fold 평균 Precision은 0.166, Recall은 0.338로 빈도기반의 Precision 0.049 보다 약 3.4배, 그리고 Recall 기준 1.5배 정도 높은 성능을 기록하였다.

2. 관세청 시스템(BRMS) 운용 평가

제안한 방법론의 실제적 평가를 위해 실험평가와는 실제 선별 운영시스템의 BRMS([그림 2] 참고)에 탑재하여 성능을 측정하였다. 운영시스템에는 본 연구에서 제안한 우범공급망 선별모델 이외에 수백개의 선별모델이 탑재되어 모델별로 우범성을 판단하여 중요적발 및 다양한 위반사항을 적발하고 있다. 모델 구축은 학습은 모델의 시스템 적용일을 기준으로 직전 6개월치(2018년 7월~2018년 12월)의 수입신고 데이터를 이용하여 모델을 매일 새로이 구축하였다. 새로이 구축된 모델들은 일단위로 시스템에 업데이트하여 적용되었으며, 2019년 1월부터 5월 17일까지 사이 약 5개월간의 운영데이터를 활용하여 결과를 평가하였다. 평가결과 중요적발률(Precision)이 기존 BRMS에서 최고 성능을 나타냈던 모델보다 10.4% 더 높게 나타났고, 운영모델 중 가장 높은 성능을 기록한 것으로 확인되었다[1].

Ⅴ. 결론

본 연구에서는 국가 관세위험을 최소화하는 일환으로 연관관계분석(Association Rule Analysis)활용하여 우범공급망 간 위험화물 선별모델 구축 방법을 제안하였다. 관세청 수입신고서 거래 데이터에서 추출한 공급망 정보와 적발정보를 이용하여 연관관계분석 방법 중 하나인 어프라이어리(Apriror) 알고리즘을 적용해 납세의무자와 해외거래처간의 우범공급망 연관규칙(Rule set)을 구축하고 연관규칙의 신뢰도(Confidence)를 활용해 위험 공급망을 선별하였다. 2년 6개월치의 데이터를 활용한 5겹 교차검증 실험결과 평균 0.166의 Precision 및 0.338의 Recall을 기록하였으며, 관세청 시스템에 적용한 실제 운영평가에서 가장 높은 중요적발률을 기록하여 10.4%(Precision)의 성능 향상을 이룬 것으로 확인되어 제안한 모델의 효용성을 증명하였다.

이로써 적발률을 제고하고 중요적발(세액) 관련 적발건 추가 추징이 가능하게끔 하고, 정상화물에 대한 업체의 불필요 검사비용 절감하도록 하였다. 또한 관세국경 위험차단에 필요한 정보와 분석환경 구축으로 다변화되고 있는 위험에 대한 과학적 선별이 가능하게 하였다. 향후연구로 공급망 정보를 그래프기반의 데이터로 모델링하여 그래프기반의 딥뉴럴 네트워크(Deep Neural Network)를 적용하여 성능향상을 도모하고자 한다.

1 운영모델명, 방식, 적발률 수치 등은 관세청 규정상 국가 대외비로 공개가 불가능하다.

* 감사의 글

연구에 도움을 주신 관세청 관세국경위험관리센터 분들에게 깊은 감사를 드립니다.

참 고 문 헌

[1] 이명구, 이은재, "4차 산업혁명기술을 적용한 관세행정 개선방안에 관한 연구," 관세학회지, 제19권, 제1호, pp.3-24, 2018.

[2] 관세청, "특송물품 수입통관 사무처리에 관한 고시," 제2017-25호, 2017.

[3] 관세청, "인공지능 기반 우범화물 검사대상 선별시스템 구축," 첨단 정보기술 활용 공공서비스 지원사업 제안요청서, 2018.

[4] 최현철, *여행자 휴대품 통관제도의 개선방안에 관한 연구,* 건국대학교, 석사학위논문, 2005.

[5] 권은주, *관세행정 위험관리제도의 효율적 운영방안에 관한 연구,* 한국해양대학교, 석사학위논문, 2014.

[6] 정분도, 김지훈, 홍미선, "전자통관시스템 활성화방안에 관한 연구-주요국의 통관단일창구 비교를 중심으로," e-비즈니스연구, 제16권, 제3호, pp.293-312, 2015.

[7] 송선욱, "국제항공화물 공급망 안전을 위한 미국의 항공화물사전검색(ACAS)에 관한 연구," 관세학회지, 제14권, 제1호, pp.133-152, 2013.

[8] 서동민, 김재수, 송정아, 박문일, "네트워크 탐색 기술

을 기반으로 한 무역 거래 위험 요소 적발 시스템 개발," 한국콘텐츠학회 2018년도 춘계 종합학술대회 논문집, pp.11-12, 2018.

[9] J. Han, M. Kamber, and J. Pei, "Data mining: concepts and techniques," The Morgan Kaufmann Series in Data Management Systems, 2011.

[10] R. Agrawal, T. Imieliński, A. Swami, R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," In Proceedings of the SIGMOD '93, 1993.

[11] 관세청, "무역통계부호," 관세청통관기획과, 2018.

[12] P. Harrington, "Machine Learning in Action," Manning Publications, 2012.

저 자 소 개

김 성 찬(Seongchan Kim) 정회원

- 2010년 2월 : KAIST 정보통신공학과 (공학석사)
- 2017년 2월 : KAIST 지식서비스공학과(공학박사)
- 2017년 2월 ~ 현재 : 한국과학기술정보연구원 연구원
- 2021년 3월 ~ 현재 : UST 데이터&HPC 과학과 겸임교수

〈관심분야〉 : 딥러닝, 기계학습, 데이터마이닝

송 사 광(Sa-Kwang Song) 정회원

- 1999년 2월 : 충남대학교 컴퓨터과학과(공학석사)
- 2011년 2월 : KAIST 전산학과(공학박사)
- 2010년 12월 ~ 현재 : 한국과학기술정보연구원 책임연구원
- 2014년 3월 ~ 현재 : UST 데이터&HPC 과학과 교수

〈관심분야〉 : 딥러닝, 빅데이터, 기계학습, 텍스트마이닝, 자연어처리, 정보검색

조 민 희(Minhee Cho) 정회원

- 2005년 2월 : 연세대학교 전산학과(이학석사)
- 2005년 4월 ~ 현재 : 한국과학기술정보연구원 선임연구원

〈관심분야〉 : 데이터마이닝, 자연어처리, 연구데이터

신 수 현(Su-Hyun Shin) 준회원

- 2017년 2월 ~ 현재 : 관세청 관세국경위험관리센터

〈관심분야〉 : 우범화물검사 선별시스템