한국빅데이터학회
제3권 제2호, 2018, pp. 19-33

수입식품 빅데이터를 이용한 부적합식품 탐지 시스템에 관한 연구

Study on Anomaly Detection Method of Improper Foods using Import Food Big data

조상구[1] · 최경현[2†]

식품안전정보원, 책임연구원[1], 한양대학교 기술경영대학원, 교수(원장)[2]

요 약

FTA체결의 증가, 식품교역 증가 및 소비자의 다양한 식품 선호도 등으로 농축수산물 및 가공식품의 수입량은 매년 증가하고 있는 추세이다. 수입식품의 안전성을 확인하는 정밀검사는 전체 수입식품건수 대비 20%정도를 차지하고 계속 증가하고 있는 반면에 정부의 수입안전관리에 필요한 예산과 인력은 그 한계점에 다다르고 있다. 수입식품 안전사고가 발생하게 되면 막대한 사회적, 경제적 손실을 야기할 수 있으므로 수입식품의 수입허용여부를 정확하게 예측하여 선제 대응하는 것은 수입안전관리의 효율성과 경제성을 획기적으로 높일 수 있게 된다. 식품분야에서는 이미 엄청난 양의 정형 데이터가 과거로부터 쌓여 왔으며 이에 대한 충분한 분석을 통한 활용은 아직은 부족한 것이 현실이다. 전체 수입건수와 중량 중에서 차지하는 가공식품의 비중은 평균 75%에 달하고 있어 식품분야에서도 빅데이터의 분석, 분석기법의 적용 등으로 다량의 데이터로부터 의미 있는 정보를 추출하는 과학적이고 자동화된 부적합탐지시스템의 연구가 절실한 상황이다 이러한 배경에서 본 연구는 기계학습분야의 다양한 부적합 예측 모형을 적용하였으며 예측 모형의 정확도를 개선시키기 위한 방편으로 새로운 파생변수의 생성을 통한 데이터 전처리 방안을 제시하였다. 또한 본 연구에서는 기계학습분야의 일반적인 기저 분류기를 적용하여 예측 모형의 성능을 비교하였으며 여러 기저 분류기 중 Gaussian Naïve Bayes예측 모형이 수입식품의 부적합을 탐지하여 예측하는 가장 좋은 성과를 보여주었다. 향후 Gaussian Naïve Bayes 예측 모형을 이용한 부적합 탐지 모형을 적용하여 수입식품의 정밀검사 비중을 낮추고 부적합률을 제고시킴으로써 수입안전관리 국가사무의 효율성과 수입통관의 신속성에 지대한 효과를 거둘 수 있으리라 기대한다.

■ 중심어 : 수입식품, Gaussian Naive Bayes 예측모형, 부적합 탐지 모형

Abstract

Owing to the increase of FTA, food trade, and versatile preferences of consumers, food import has increased at tremendous rate every year. While the inspection check of imported food accounts for about 20% of the total food import, the budget and manpower necessary for the government's import inspection control is reaching its limit. The sudden import food accidents can cause enormous social and economic losses. Therefore, predictive system to forecast the compliance of food import with its preemptive measures will greatly improve the efficiency and effectiveness of import safety control management. There has already been a huge data accumulated from the past. The processed foods account for 75% of the total food import in the import food sector. The analysis of big data and the application of analytical techniques are also used to extract meaningful information from a large amount of data. Unfortunately, not many studies have been done regarding analyzing the import food and its implication with understanding the big data of food import.

2018년 11월 14일 접수; 2018년 11월 15일 수정본 접수; 2018년 12월 31일 게재 확정.
† 교신저자 ghchoi@hanyang.ac.kr

In this context, this study applied a variety of classification algorithms in the field of machine learning and suggested a data preprocessing method through the generation of new derivative variables to improve the accuracy of the model. In addition, the present study compared the performance of the predictive classification algorithms with the general base classifier. The Gaussian Naïve Bayes prediction model among various base classifiers showed the best performance to detect and predict the nonconformity of imported food. In the future, it is expected that the application of the abnormality detection model using the Gaussian Naïve Bayes. The predictive model will reduce the burdens of the inspection of import food and increase the non-conformity rate, which will have a great effect on the efficiency of the food import safety control and the speed of import customs clearance.

■ Keyword : Import food, Gaussian Naïve Bayes classifier, Predictive system

Ⅰ. 서 론

1995년 WTO(World Trade Organization)출범 이후 국간 간 자유무역협정(FTA) 체결 증가로 인해 수입식품은 지속적으로 증가하고 있는 추세이다. 최근 5년간 농축수산물 및 가공식품의 수입건수와 수입중량 증가율은 각 각 7.5%, 6.0%로 매년 증가하고 있다. 2017년도에는 총 167개국으로부터 625,443건을 수입하였고 이중 가공식품은 총 440,974건으로 전체 수입신고의 71%를 차지하고 있다. 따라서 불량식품, 방사능 오염, GMO식품 등의 식품안전사고 방지를 위한 수입통관 단계의 안전관리는 더욱 중요해질 것으로 예상된다. 수입식품 등에 대한 통관단계 검사는 서류검사, 현장검사, 정밀검사 등으로 구분하여 이루어지고 있다. 2017년도의 정밀검사는 140,853건으로 전체 수입건수의 22.52%이다. 정밀검사는 수입식품의 표본을 직접 채취하여 실험실에서 이화학검사를 실시하는 등 시간과 비용의 부담이 있어 식품의약품안전처의 한정된 자원으로 검사를 확대하는 것은 어려운 현실이다. 국민건강에 위험을 끼칠 수 있는 수입식품을 정밀검사결과를 사전에 예측하여 식품안전사고를 방지하는 것이 필요하다. 사회적, 경제적 손실을 야기할 수 있는 불량대출, 카드사고 등과 같이 수입식품의 부적합 여부를 사전에 정확하게 예측하여 사전 대응하는 것이 대단히 중요한 의사결정 문제 중 하나이다. 수입식품의 부적합 예측 모형의 성과는 불량 수입식품을 우량으로 판정하는 오류를 줄이는 것과 함께 우량 수입식품에 대한 적합판정을 높여 수입통관의 절차를 신속하게 하여 업무의 효율성을 높이는 것에 달려 있다.

아쉽게도 지금까지 수입식품의 현황에 대한 데이터 분석과 기계학습 예측 모형의 적용에 관한 연구가 국내외에서 거의 없는 실정이다. 본 연구는 수입식품 데이터를 정제하고 부적합 예측 모형의 적용을 통해 사전 예측 모형의 성과를 높이고자 한다. 수입식품 데이터의 부적합비율은 1.01%로 심한 불균형 데이터 형태를 갖고 있다. 부도예측, 카드사고, 이탈고객, 스팸메일, 의료진단 등의 분야에서 나타나는 불균형 데이터에 관한 예측모형의 정확도(Accuracy)는 다수범주에 속하는 패턴에 대한 정확도로 소수범주에 대한 예측성과의 척도로는 적당하지 않다. 불균형 문제를 해결하기 위한 대표적인 방법으로는 샘플링을 이용한 방법과 오 분류를 조정하는 방법이 있다(Kim and Hong, 2014). 의사결정나무 모형, Random Forest 및 앙상블모형(Random Forest, Gradient Boosted Trees 등) 등은 나무 형태를 갖는 계층적 구조의 알고리즘으로 다수와 소수 범주의 패턴을 구분하여 예측하는 데 적합하다(Galar, et al., 2012). 부적합 수입식품 사전예측 기법은 부적합과 적합을 사전에

`구분하여 예측하는 지도학습 기법(The supervised Method), 수입식품부문 전문가집단의 자문을 통해 이상치를 구분하여 예측하는 준지도학습 기법(The Semi-supervised Method) 및 군집분석(Clustering Analysis) 등이 있다(Pai, et al., 2014). 최근 데이터 마이닝, 기계학습 분야에서 여러 기저 분류기와 관련된 연구는 의학, 금융, 마케팅 부문에서 다양하게 이루어져 왔다(Koufakou and Georgiopoulos, 2010, Otey, et al., 2006).

본 연구에서는 수입식품 데이터의 특성을 고려하여 기존의 다양한 기저분류기의 적용과 더불어 수입통관업무 전문가와 FGI(Focused Group Interview)를 통해 데이터를 정제하고 최적의 파생변수를 생성하여 예측모형의 성과를 분석하였다. 데이터의 불균형문제는 예측 모형의 성능을 저하시키는 요인으로 작용하고 있으며 이를 해결하기 위해 금융, 의료 및 마케팅 분야의 부도예측 및 신용불량자 추출, 희귀한 질병을 가진 환자의 진단 및 이탈고객 방지 등 많은 연구가 이루어지고 있다(Lee and Kwon, 2013; Kang et al., 2004; Mac Namee, et al., 2002). 데이터 불균형의 문제를 해결하기 위한 방법으로 데이터 축소(data reduction)기법, 사례선택기법 등이 있다. 데이터 축소 기법은 데이터 전처리(data reprocessing)기법의 하나로 대량의 데이터를 보다 가장 대표성이 있는 핵심적인 데이터를 선정하여 원 데이터의 크기를 축소시키는 것이다. 사례선택기법은 원 데이터를 사용하는 것보다 축소된 데이터를 사용함으로써 원 데이터를 사용하는 것보다 유사하거나 더 나은 예측모형을 도출하는 것이다(García, et al., 2012). Derrac, et al.(2012)은 사례선택을 선택하는 방법에 따라 Wrapper기법과 Filter기법으로 구분하였고 이의 실증을 위해 k-nearest neighbors 기법을 사용하여 학습용 데이터(training data)를 크게 축소하면서도 모형의 성과를 높일 수 있다는 것을 보여 주었다. 불균형 문제를 해결하기 위해 샘플링을 이용한 방법과 오 분류를 조정하는 방법이 대표적이다. 다수 범주집단의 데이터를 임의로 샘플링하는 under sampling과 소수 범주집단의 데이터를 반복적으로 복사하여 샘플링하는 over sampling방법이 있다(Ganganwar, 2012; Liu, et al., 2006). 과거에 적용되었던 사례와 그 결과를 참조하여 새로운 사례에 적용하는 사례기반추론과 같이 특화된 지식을 활용하여 불균형데이터문제를 해결하는 방법도 있다(Allen, 1994). 이 이외에도 클러스터링 방법을 통한 데이터 전처리 기법, 유전자 알고리즘을 통해 결합적으로 데이터의 불균형 문제를 해결하고자 하이브리드 모형을 적용하여 소수 범주 데이터의 패턴을 찾아내고자 하였다(Hwang, et al., 2007).

정보기술의 빠른 발전, 빅데이터의 등장, 데이터 분석 기법의 고도화 등으로 인해 대량의 데이터를 처리하고 분석하는 연구가 다양한 분야에서 다 학제적으로 이루어지고 있다(Wu, et al., 2008). 논문의 구성은 다음과 같다. 2장에서는 수입식품관리 현황 및 부적합식품 예측을 위한 현재의 방안 등에 대해 기술하였다. 3장에서는 실증분석을 위해 수집한 수입식품 데이터의 특성을 고려하여 FGI결과를 기반으로 새로운 파생변수를 생성하였다. 4장에서는 실증분석을 위해 수집한 수입식품신고 현황에 대한 기초통계를 실시하였고 기계학습의 다양한 기저 분류기를 적용하여 민감도(sensitivity), 특이도(specificity), 수신자조작특성곡선(ROC, Receiver Operating Characteristic Curve) 등을 산출하여 예측모형의 성과를 평가하였다. 5장에서는 연구의 결과와 시사점을 정리하였다.

Ⅱ. 수입식품관리 현황

국내 소비자의 식품 선호도 확대에 따라 수입

〈표 1〉 Increasing trend of food import from 2012 to 2016

Year	Import entry	Growth rate (%)	Weight(Ton)	Growth rate (%)	Price($)	Growth rate (%)
2012	474,648	+ 0.3	15,837,144	+ 0.1	21,333,762,434	+ 0.8
2013	494,242	+ 4.1	15,541,310	(- 1.9)	21,551,768,473	+ 1.0
2014	554,177	+ 12.1	16,358,300	+ 5.3	23,111,675,025	+ 7.2
2015	598,082	+ 7.9	17,064,298	+ 4.3	23,294,688,821	+ 0.8
2016	625,443	+ 4.6	17,260,883	+ 1.2	23,437,592,852	+ 0.6

〈표 2〉 import inspection status by the type of inspection from 2014 to 2016

Year	Number of Import entry	Screening	Field test	Laboratory test	Reject rate
2014	554,177	343,483 (62.0%)	97,836 (17.7%)	112,858 (20.4%)	0.84%
2015	598,082	373,570 (62.5%)	101,422 (17.0%)	123,090 (20.6%)	0.96%
2016	625,443	377,916 (60.4%)	106674 (17.1%)	140,853 (22.5%)	0.84%

식품 유형 및 수출·입 국가가 다변화되고 있는 추세이다. ‘12년도부터 최근 5년간 수입식품의 변화는 <표 1>과 같이 ‘16년도에는 '12년도 대비하여 수입건수는 31.7%, 중량은 8.9%, 금액은 연간 9.8% 증가하였다.

수입식품관리는 수입국의 정치, 경제, 사회, 문화적 특성 등을 고려하여 여러 가지 정책 수단의 최적화를 통해 효과적이고 효율적으로 이루어지는 의사결정 과정이다. 수입식품 안전관리 강화의 필요성 대두에 따라 2015년 2월에 「수입식품안전관리 특별법」을 제정(2015.2.3)하여 수입 전(前), 통관 및 유통단계 등 3중의 ‘수입식품 안전관리망’을 구축하고 수입업소가 스스로 안전성을 책임지는 정책을 추진하고 있다 (Chang and Lee, 2016). 매년 증가하는 수입검사에 대한 효율적인 업무처리 방안으로 수입 통관 단계에서 부적합식품을 사전에 예측하는 위험분석기반(Risk Analysis based) 자동화 시스템의 구현이 필요하다. 정부는 국가, 식품유형, 제조업소 및 수입업소 등의 과거이력, 검사결과 등의 분석을 통해 수입식품 등급을 분류하는 수입검사시스템을 구축하여 운영하고 있다.

국내로 수입되는 식품의 과거 부적합 내역, 위해 정보, 수입업소 및 제조업소 등의 과거 행정위반 이력 등을 종합 판단하여 집중관리 대상 및 항목을 실시간으로 자동 추출하고 있다. 수입식품은 검사대상에 따라 서류검사, 관능검사, 정밀검사 등으로 나누어진다. 수입식품 검사유형별에서 정밀검사가 차지하는 비중은 <표 2>에서 보면 2014년부터 2016년까지 평균 21.19%를 차지하고 있다. 정밀검사 결과 부적합 판정율은 2013년도이후 최근 3년동안 0.83%, 0.96%, 0.84%를 나타내고 있다. 수입식품의 안전사고를 방지하기 위해서 정부는 정밀검사 비중을 늘리거나 부적합판정비율을 높일 수 있는 방안을 강구하여야 한다. 본 연구에서는 정밀검사 결과에 대한 사전예측을 통해 정밀검사의 부적합 판정율을 높이고자 한다.

미국 FDA(Food and Drug Agency)는 PREDICT(Predictive Risk-based Evaluation for Dynamic Import Compliance Targeting)시스템을 활용하여 내부 자료와 외부의 다양한 채널을 통

해 얻은 자료를 기반으로 규칙기반 전문가시스템(Rule based Expert System)과 데이터 마이닝을 통해 과거 자료를 분석하여 수입식품의 위험등급을 산출하여 부적합이 예상되는 식품을 사전에 예측하고 있다. 수입식품에 대한 부적합 여부를 사전에 예측하게 되면 정밀검사 비율을 줄여 통관지연으로 인한 민원을 줄일 수 있으며, 효율적인 검사업무로 수입통관담당 검사소 담당인력의 업무 부담 감소와 국가 물류비용 감소 등으로 대외경쟁력을 강화할 수 있다.

Ⅲ. 데이터 수집 및 정제

3.1 데이터 수집 및 구성

본 연구의 데이터는 2014년부터 2018년6월말까지 총 수입신고 건을 대상으로 하여 이 중에서 농축수산물을 제외한 식품(가공식품, 건강기능식품, 첨가물, 기구 또는 용기·포장 등)만을 대상으로 하였다. 식품 관련 수입신고 현황자료는 2014년 393,216건, 2015년 426,272건, 2016년 440,974건, 2017년 474,477건 및 2018년 6월말 254,632건 등 총 1,989,571건의 데이터마트를 1차적으로 구축하였다. 정밀검사의 판정결과가 있는 경우만을 데이터로 포함시키기 위해 수입업소가 자진 취하한 경우, 수입검사원에 의해 반려된 경우, 부분 부적합으로 인한 보류 등은 수입식품의 정밀검사 결과의 판정에 관련이 없어 데이터구성에서 제외하였다. 최종적으로 식품만으로 구성된 총 392,454건의 데이터마트를 마련하였다. <표 3>을 보면 식품유형별로는 가공식품의 수입건수가 304,442건(77.57%), 수입중량은 207천만ton(94.12%) 및 수입관세가격은 44억8천만 USD(75.83%)를 나타내고 있다. 수입부적합비율기준으로는 기구, 포장·용기는 2.47%, 건강기능식품은 1.1%, 가공식품은 0.98%, 식품첨가물은 0.55% 순으로 나타나고 있으며 전체 수입식품의 부적합율은 4,279건으로 전체 수입건수 392,454건 대비 1.09%를 차지하고 있다.

정밀검사 결과 '적합'인 경우는 388,175건(98.91%), '부적합'인 경우는 4,279건(1.09%)으로 불균형데이터의 형태를 보여주고 있다. 최종 데이터는 총 34개의 변수로 구성되어 있었으나 FGI(Focused Group Interview)를 통해 사업자 내부정보, 고유코드 등 다수의 변수를 제거 한 후 최종적으로 <표 4>와 같이 수입용도, 수입업소, 식품유형, 국가, 수입가격, 수입중량 및 정밀검사 결과 등 7개의 변수를 선정하였다. 데이터의 계절효과(Seasonal effects)는 나타나지 않았으며 제조국가(원산지)와 수출국가의 차이로 인한 정밀검사 결과는 거의 상관관계가 없는 것으로 나타나 수입년도·월, 수출국 등의 변수는 분석대상에서 제외하였다.

〈표 3〉 Import inspection types of food import in 2016

Unit : import entry, 10million ton, 10million USD

	Import entry (%)	Reject to import (%)	Import weight (%)	Import price (%)
Processed food	304,442 (77.6%)	2,989 (0.98%)	207.32 (94.12%)	448.02 (75.83%)
Dietary food	41,715 (10.6%)	463 (1.11%)	1.09 (10.49%)	56.94 (19.64%)
Food container	29,814 (17.6%)	737 (2.47%)	5.81 (12.64%)	47.58 (18.05%)
Food additives	16,483 (14.2%)	90 (0.55%)	6.06 (12.75%)	38.30 (16.48%)
total	392,454	4,279 (1.09%)	220.27	590.84

〈표 4〉 Variable list

Variable Name	Variable Description	Data Type	Categories
Import purpose	Sales, process input, R&D, etc	Categorical	10
Importer	Company to import food items	Categorical	15,473
Food type	Beef, fish, Agricultural product, processed food, etc	Categorical	1,204
Origin of country	the country from which a food originally comes	Categorical	135
Import price	Custom price of import food	Numerical	NA
Import weight (KG)	Weight (KG) of import food	Numerical	NA
Result of Lab test	Lab inspection test results to reject or not to import	Integer	2

3.2 데이터 정제 및 파생변수

수입식품 데이터 정제단계에서 데이터 수집으로부터 도출한 변수 중 중복되거나 수입식품의 부적합 판정 결정에 영향을 미치지 못하는 변수는 제거하였다. 국제적인 식품안전 정책의 방향은 위해요소(hazard)에 대한 관리뿐만 아니라 식품유형 및 국가 등에 대한 위험(risk)을 기초로 가장 효과적이고 경제적인 수입식품 검사 방법을 권장하고 있다(Hoffmann, 2005). 국내 수입식품검사는 국가, 식품유형 및 수입업소 등의 개별적인 위해요소(hazard)와 위험(risk)를 고려하여 수입식품의 위험을 평가하고 있다. 본 연구에서는 수입식품 데이터를 대상으로 2015년부터 국가, 수입업소 및 식품유형별로 수입건수, 수입중량 및 가격, 부적합건수, 부적합률 등을 추가적인 파생변수로 생성하였다. 이외에도 국가와 식품유형, 국가와 수입업소 및 식품유형과 수입업소별 수입중량, 수입가격, 수입건수 및 부적합률 등 수입현황관련 각 종 비율자료를 생성하여 수입식품의 부적합을 예측하기 위한 변수로 사용하였다. 이는 재무 및 회계부문 등에서 보험사기, 이상적인 금융거래, 부도예측 등의 불균형데이터에서 적용되어온 초기의 예측 모형의 단일변량 혹은 다중변량의 변수 기반의 통계 모형을 적극 반영한 것이다(Beaver, 1996; Ohlson, 1980).

수입식품의 식품유형과 수입국 등에 대한 과거의 부적합 이력을 근거로 수입검사관리 업무가 이루어지고 있으며 식품안전관리는 과거에 발생했던 식품안전관련 사건 및 사고에 대한 분석이 우선시 되는 것과 동일한 맥락을 갖고 있다(Park, et al., 2017). 수입식품 부적합 판정에 관련이 있다고 판단되는 수입검사 관련 비율의 파생변수를 FGI를 통해 총 18개의 파생변수를 <표 5>와 같이 생성하였다. 국가, 식품유형 및 수입업소를 기준으로 수입신청중량, 가격 및 부적합률을 생성하였고 국가와 식품유형, 국가와 수입업소 및 수입업소와 식품유형을 그룹으로 하여 각각의 중량, 가격 및 부적합률 등을 파생변수로 생성하였다. 연속형 입력변수인 수입신고 중량과 가격은 데이터의 정규화를 하여 입력변수 값에 선형변형을 적용하여 자료의 분포를 평균 0, 분산 1이 되도록 하였다.

수입용도, 수입업소, 국가, 식품유형, 수입신고가격 및 수입신고중량 등 기본 입력변수인 6개와 파생변수 18개를 합하여 총 24개의 변수를 생성하였다. 입력변수간 다중공선성(multicollinearity)을 판단하기 위해 분산팽창계수(Variance Inflation Factor, VIF)가 높은 국가와 수입업소별 중량(Import weight_groupby(country.importer), VIF=29.44), 용도(VIF=22.46), 국가별 가격가격(Import price_country of origin, VIF=13.56), 수입업소

〈표 5〉 derivative variable list

Derivative variable name	Data type	Number of uniqueness
Import weight_country of origin	Float	135
Import price_country of origin	Float	135
Reject rate_country of origin	Float	135
Import weight_importer	Float	15,473
Import price__importer	Float	15,473
Reject rate__importer	Float	15,473
Import weight_food type	Float	1,204
Import price__food type	Float	1,204
Reject rate__food type	Float	1,204
Import weight_groupby(country.importer)	Float	29,626
Import price_groupby(country.importer)	Float	29,626
Reject rate_groupby(country.importer)	Float	29,626
Import weight_groupby(origin.foodtype)	Float	7,804
Import price_groupby(origin.foodtype)	Float	7,804
Reject rate_groupby(origin.foodtype)	Float	7,804
Import weight_groupby(importer.foodtype)	Float	52,643
Import price_groupby(importer.foodtype)	Float	52,643
Reject rate_groupby(importer.foodtype)	Float	52,643

와 식품유형별 수입중량(Import weight_groupby(importer.foodtype), VIF=13.28), 식품유형별 가격(Import price__food type, VIF=7.29), 수입업소별 중량(Import weight_importer, VIF=5.37), 국가(VIF=5.26) 등의 7개 입력변수는 예측모형에서 제외하였다((Thompson, et al., 2017). 실험에 사용한 최종 입력변수는 로짓회귀분석을 이용하여 부적합 판정결과 변수와의 임계값 p-value값이 0.01보다 적은 입력변수는 모형에 반영하였고, 0.05보다 큰 입력변수는 모형에서 제외하여 최종 6개의 입력변수를 선정하였다. 최종 입력변수는 국가와 식품유형별 수입부적합률, 국가와 수입업소별 수입부적합률 및 식품유형과 수입업소별 수입부적합률 등이다.

Ⅳ. 실증분석

불균형데이터를 비율을 비슷하게 하기 위해 샘플링방법이나 오분류 비용을 통해 전체 예측모형의 성과를 향상(Barandela, et al., 2003)시키고자 하였으나 예측모형의 성과가 향상되는 것이 아주 적었다. 실증분석은 '기초 통계분석', '예측모델링', 모형'평가'순으로 작성하였다.

4.1 기초 통계분석

수입식품 데이터에 대한 이해도를 높이고 수입부적합 및 적합 판정결과에 미치는 변수의 변별력을 파악하는 목적으로 수행하였다. 총 135

개의 수입국가 중 중량기준 상위 20개국(중국, 호주, 중국, 러시아연방, 브라질, 베트남, 필리핀 등)의 수입량이 전체 수입량의 92.6%를 차지하고 총 수입식품 총 254개 식품유형 중에 중 중량기준 상위 20개 식품유형(정제, 가공용 식품원료, 배추김치, 과·채가공품, 서류·당류·곡류·두류 가공품 등)의 수입량이 전체 수입량의 83.6%를 차지하고 있다. 국가와 식품유형의 부적합률이 높은 고위험군 국가와 식품유형의 데이터를 추출하여 분석하였으나 예측모형의 성과를 기대할 수는 없었다. 개별 변수들의 평균 및 표준편차를 비교한 후 변수 간 상관관계를 분석하였다. <표 6> 은 부적합과 적합판정 결과에 대한 입력변수들의 평균 값이다. 부적합과 적합판정 결과가 나온 수입신청가격의 평균 값은 각각 10,250USD와 15,108USD로 부적합판정의 수입신고 평균 값이 적합판정에 비해 67.85%적다고 해석되며 대부분의 입력변수의 부적합과 적합의 비율차이가 있는 것으로 나타난다. 제조국과 식품유형과 관련된 입력변수의 비율차이가 다른 입력변수 평균 값의 비율보다 현저히 차이가 있는 것으로 나타나 중요한 변수라는 것을 알 수 있다.

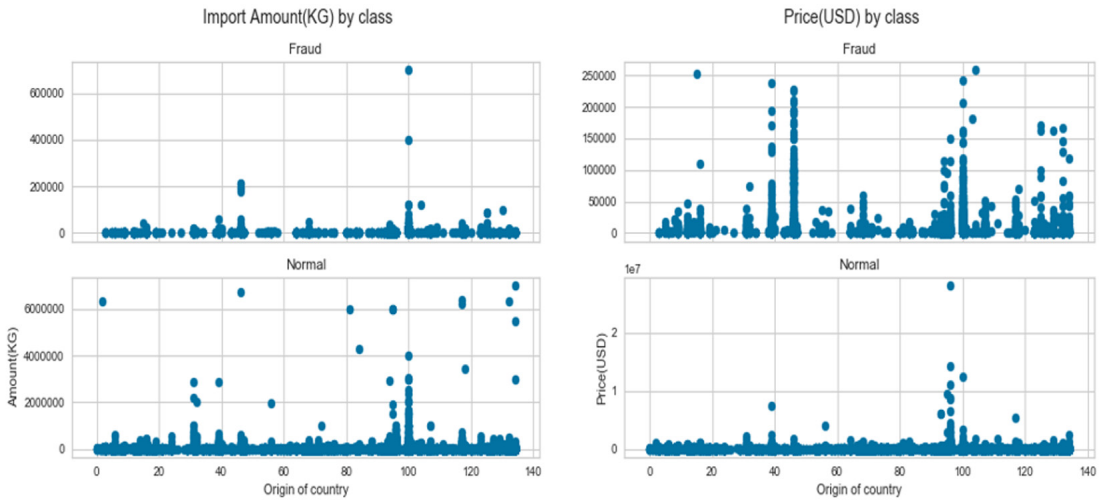
<그림 1>, <그림 2>는 국가 및 수입업소업별로 정밀검사 결과별 수입 중량과 가격을 보여주고 있다. 정밀검사 결과 부적합판정을 받은 국가가 수입 중량과 수입가격 등의 상대적 규모가 적다. 부적합판정을 받은 수입업체의 경우도 비슷한 형태를 보이고 있다.

4.2 예측 모델링 및 평가

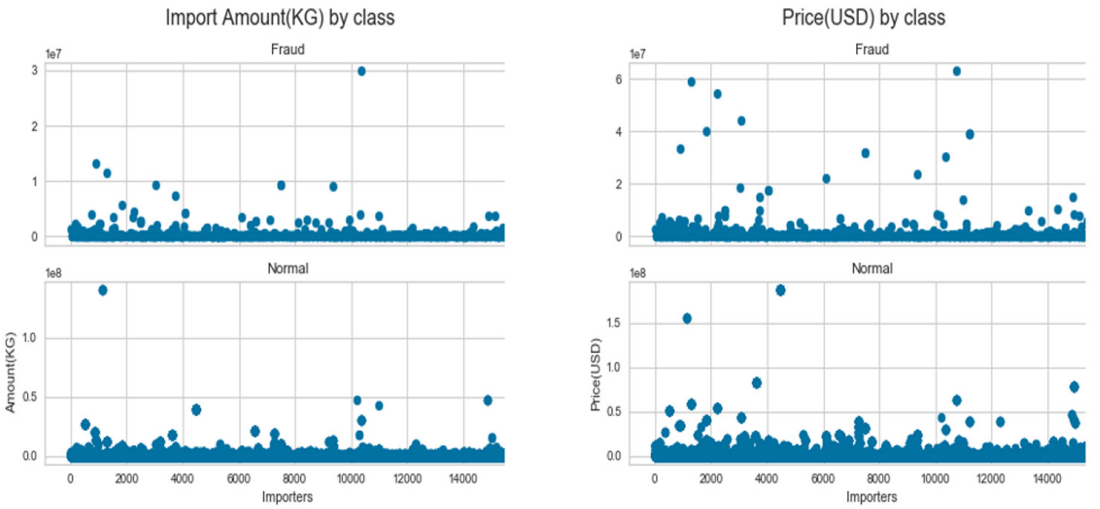
수입식품 검사 시 국가와 식품유형을 개별적으로 구분하여 검사를 하지 않고 국가별 식품유형 혹은 수입업소별 식품유형 등과 같이 국가, 수입업소 및 식품유형 등을 동시에 고려하여 수입검사관리업무를 수행하고 있다.「수입식품안전관리 특별법」 제 21조(수입검사 등)에 의하면 수입검사를 할 때에는 수입식품 등의 검사이력, 국내외 식품안전정보 등에 따라 수입식품 등을 구분하여 차등 검사할 수 있다라고 규정하고 있다. 수입식품의 정밀검사를 무작위 표본검사로

〈Table 6〉 Comparison of import statistics between rejected and accepted category

		Reject	Accept	A/B (%)
Average Import price (USD)		10,250	15,108	67.85
Average Import weight (ton)		3,008	5,641	53.33
Country of origin	Average import entry line	32.470	57,664	56.31
	Average Import price (USD)	562,361,405	844,996,420	66.55
	Average Import weight (ton)	258,741,553	314,770,452	82.20
	Rate of rejection to import	2.36%	1.08%	
Food type	Average import entry line	4,213	8,035	52.43
	Average Import price (USD)	74,142,366	92,638,074	80.03
	Average Import weight (ton)	25,381,236	31,218,051	81.30
	Rate of rejection to import	4.91	1.05%	
Importer	Average import entry line	697%	2,923	23.85
	Average Import price (USD)	16,303,139	32,453,290	50.24
	Average Import weight (ton)	4,760,449	11,241,863	42.35
	Rate of rejection to import	1.62%	0.92%	



〈그림 1〉 Import amount and price value by country among reject and accepted ones



〈그림 2〉 Import amount and price value by importer among reject and accepted ones

수행할 경우 부적합 이력이 있는 국가별 식품유형, 수입업소 및 제조업소 등에 따라 차등 적용하여 매년 수행하여 오고 있다. 본 연구의 입력변수인 국가, 수입업소 및 식품유형별 부적합률, 국가와 수입업소별, 식품유형과 수입업소 및 국가와 식품유형 등의 부적합률은 전국 16개 수입검사소에서 관리하고 있는 실정을 충분히 반영한 것으로 판단된다. 예측모형 적용대상 수입식품데이터는 학습용 데이터와 모형검증을 위한 검증용 데이터로 분할하여 실험을 수행하였다.

테스트 방법은 10겹 교차검증(10-fold cross validation)방법으로 수행하였다. 수입 용도, 수입업체, 식품유형 및 국가 등과 같은 범주형데이타의 인코딩(Encoding)방식은 Label Encoder 방식을 적용하였다. 데이터의 오버플로우(overflow)나 언더플로우(underflow)를 방지하기

위해 실수형태의 입력변수 값은 정규화하여 예측모형의 계수추정과 입력데이터의 변환을 동시에 실행하였다. 수입부적합 예측모형은 지도학습알고리즘으로 Decision Tree, K-nearest neighbors, Naïve Bayes, Random Forrest, Ensemble 구분자를 사용하여 부적합과 적합 판정결과를 예측하였다. 범주형 자료의 이상치 관측된 객체간의 유클리드 거리를 계산하는 방법인 거리척도 기반 알고리즘, 연관성 규칙기반 알고리즘(Said, et al., 2011; Otey, et al., 2006), 군집기반방법(Cao, et al., 2013) 및 밀도기반 기법(Zhao, et al., 2014) 등 수입식품데이터의 불균형데이터의 특성을 고려하여 DBSCAN, Isolation Forest 및 Local Outlier Factor 등의 머신러닝 알고리즘을 적용하였다.

각각 예측모형의 성과는 AUROC(Area Under ROC Curve)를 기준으로 비교하였다(Fawcett, 2006). 예측모형의 수신자조작특성곡선(ROC, Receiver Operating Characteristic Curve)은 진짜 부적합식품 중에서 부적합을 얼마나 잘 식별하는 지를 나타내는 민감도(Sensitivity)와 진짜 적합식품 중에서 예측모형의 방법이 적합식품을 얼마나 잘 골라내는지를 나타내는 특이도(Specify)를 보여주고 있다. 본 연구의 목적은 실제로 부적합식품 중 부적합 사전예측이 있다고 가려낼 확률과 실제 적합식품 중에서 적합하다는 사전예측을 할 확률을 가장 크게 줄일 수 있는 예측모형을 찾는 것이다. 따라서 수입식품의 정밀검사 결과를 부적합과 적합의 이진분류를 위한 부적합 예측 모형의 성과는 AUROC의 척도로 판단하였다.

4.3 실험결과

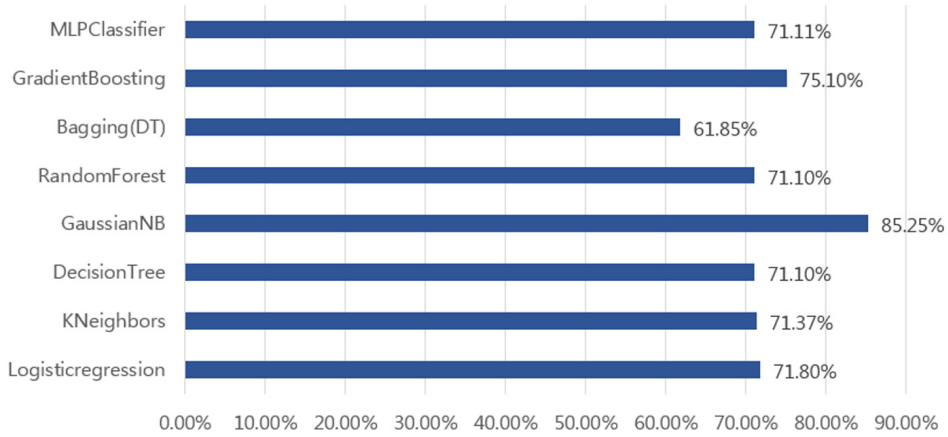
수입식품의 부적합 건수는 4,279건으로 전체 수입건수 392,454건 대비 1.09%를 차지하는 불균형데이터 형태를 나타내고 있다. 이러한 극심한 불균형 데이터의 대한 예측 모형은 적합으로 판정된 다수의 범주로 패턴분류를 많이 하게 되어 부적합으로 판정된 데이터인 소수범주는 무시되거나 다수의 범주로 취급하게 되는 경향이 있다(Jo and Japkowicz, 2004, , Kang, et al., 2004; Weiss, 2003). 여러 분류기들의 수신자조작특성곡선은 그 성과를 도식화하고 비교하기에 유용하며 ROC커브의 아래의 면적을 의미하는 AUROC(area under the ROC curve)는 여러 기저 분류기들의 성과를 평가하는데 적합하다(Jin and Ling, 2005; Cortes and Mohri, 2004).

예측 모형의 초기 입력변수는 파생변수를 포함시키지 않고 수입용도, 수입업소, 식품유형, 국가, 수입가격, 수입중량 등 6개의 입력변수로 분류형(Classification type)과 군집형(Cluster type) 머신러닝알고리즘을 적용하였다. 그 결과 의사결정나무(Decision Tree)기법의 AUC는 0.56으로 가장 높았고 다른 예측모형은 AUC가 평균 0.5수준이었다. 군집형 머신러닝알고리즘 적용결과 Isolation Forest예측모형의 AUC가 0.64이고 Local Outlier Factor와 DBSCAN 등의 AUC는 0.5수준으로 아주 낮은 recall 점수를 보여주었다. 수입식품의 불균형데이터는 적합과 부적합의 비율이 현저히 차이가 나기 때문에 예측모형의 AUROC의 수치(Fawcett, 2006)가 낮게 나타나게 경향이 있어 모형의 성능을 저하시켜 이에 대한 개선이 필요하였다. 예측모형의 성과를 높이기 위해 본 연구에서는 최적의 파생변수를 선정한 후 지도학습의 적합과 부적합을 구분하는 분류기법의 머신러닝모형만을 적용하였다.

기계학습분야의 다양한 예측 모형을 적용한 결과 Gaussian Naïve Bayes예측모형이 로지스틱 회귀분석, 의사결정나무, K-nearest neighbors, Gradient Boosting 등의 예측모형을 적용한 결과보다 <그림 3>과 같이 AUROC값이 가장 높게 나타나 예측 모형의 성과가 가장 높은 것으로

나타났다. <표 7>에서 나타나듯이 Gaussian Naïve Bayes Gaussian Naïve Bayes 모형은 실제 부적합 식품에서 부적합을 골라내는 확률인 re-call 값이 0.734으로 다른 예측모형보다 월등히 높은 반면 모형의 정확도(Accuracy)는 96.9%로 다른 예측모형과 비교하여 낮은 수준을 보임을 알 수 있다.

결론적으로 Gaussian Naïve Bayes예측모형은 부적합으로 예측하였을 경우 맞을 확률(정밀도, precision)은 21.8%, 실제로 부적합판정인 데이터를 부적합이라고 판정할 확률(recall)은 73.4%이며 AUROC는 0.853으로 다른 예측모형에 비



〈그림 3〉 Model Prediction results (AUC)

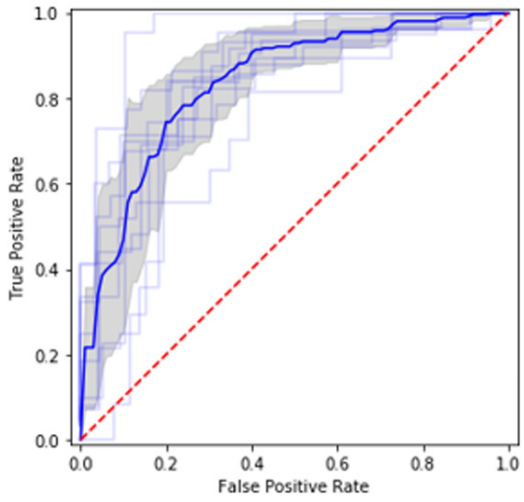
〈표 7〉 Machine learning classifiers' confusion matrix and model performance metric

Classifiers		Recall	Precision	Accuracy
Logistic Regression	Accept	0.994	0.998	0.992
	Reject	0.438	0.723	
KNeighbors	Accept	0.998	0.994	0.992
	Reject	0.429	0.740	
Decision Tree	Accept	0.993	0.998	0.992
	Reject	0.374	0.719	
Gaussian Naïve Bayes	Accept	0.971	0.997	0.969
	Reject	0.734	0.218	
Random Forest	Accept	0.993	0.998	0.991
	Reject	0.677	0.363	
Bagging (DT)	Accept	0.998	0.992	0.990
	Reject	0.239	0.582	
Gradient Boosting	Accept	0.996	0.995	0.991
	Reject	0.506	0.590	
MLP Classifier	Accept	0.998	0.994	0.992
	Reject	0.424	0.754	

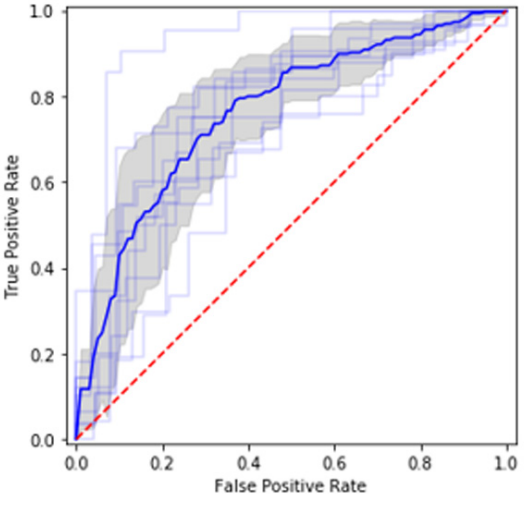
해 훨씬 높은 수치가 나타났다.

Logistic Regression 예측 모형의 경우 적합이라고 예측했을 때 실제 적합일 확률이 72.3%이며 부적합이라고 예측했을 때 실제 부적합일 확률은 43.8%이다. GNB예측모형은 부적합이라고 예측하여 실제로 맞춘 수입신고 건이 73.4%인 반면 Logistic Regression예측모형은 43.8%로 낮아 결과적으로는 AUROC가 낮아 예측 모형의 성능이 떨어진다. Gradient Boosting예측모형은 AUROC가 0.751로 2번째로 높은 성과 수치를 보여주고 있다.

부적합식품의 판정에 사전예측모형 유효성의 평가변수로 예측모형의 민감도(sensitivity), 특이도(specificity), 수신자조작특성곡선(ROC, Receiver Operating Characteristic Curve) 등을 산출할 수 있다. <그림 4>, <그림 5>는 10겹 교차검증(10-fold cross validation)방법으로 Gaussian Naïve Bayes과 Gredient Boosting예측모형을 적용한 ROC Curve를 보여주고 있다. Y축(True Positive rate)의 민감도(sensitivity)는 실제로 부적합식품 중 부적합 사전예측이 있다고 가려낼 확률, X축(False Positive rate)의 특이도(specify)는 실제 적합식품 중에서 적합하다는 사전예측을 할 확률에서 1을 차감한 값이다. 수신자조작특성곡선을 사용하여 민감도와 허위양성률(1-특이도)을 표현한 그래프로 AUROC는 아래 면적으로 예측모형의 정확도를 의미한다. <그림 4>, <그림 5>의 우상향 45도 직선으로 표시된 임의 추측 모형의 경우 직선 아래의 면적이 0.5이므로 AUROC의 값은 0.5가 되면 직선의 왼쪽 위에 ROC 커브가 위치할 경우 0.5보다 보다 큰 값을 갖게 되며 반대의 경우는 0.5보다 작은 값을 갖게 된다. AUC값은 0과 1사이의 값을 갖게 되며 값이 클수록 좋은 모형이다(Min, 2014).



〈그림 4〉 ROC curves (Gaussian NB)



〈그림 5〉 ROC curves (Gradient Boosting)

<표 8>은 국가, 식품유형과 수입업소 등을 각각 2개의 변수를 그룹핑한 데이터를 대상으로 Gaussian Naïve Bayes 머신러닝을 적용한 결과이다. 데이터의 축소로 세가지 경우 모두 AUROC(Area Under ROC Curve)가 0.95이다. Gaussian Naïve Bayes예측모형을 적용하여 부적합이 예측되는 정밀검사 조사 건수가 많을 경우에 한정된 시간과 인력의 제약조건에서 단순 참

〈표 8〉 Gaussian NB model applied to three sub-group data

Gaussian Naïve Bayes algorithm	Predict	Accept to import	Reject to import
Country of origin & food type (7,804 items)	Accept	2,885 (TP)	5 (FP)
	Reject	127 (FN)	105 (TN)
Country of origin & importer (29,296 items)	Accept	11,200 (TP)	19 (FP)
	Reject	373 (FN)	259 (TN)
Importer & food type (52,643 items)	Accept	19,766 (TP)	724 (FP)
	Reject	43 (FN)	525 (TN)

조할 수 있는 보조 지표로 활용이 가능하리라 판단된다.

Ⅴ. 결론 및 시사점

빅데이터와 기계학습 분야의 다양한 분석기법을 활용하여 수입식품 빅데이터의 특정패턴을 인식하고 분석해 정밀검사 결과를 직접 실시하기 전에 부적합식품의 탐지 예측을 수행하였다. 본 연구의 목적은 부적합이 우려되는 수입식품을 사전에 예측하여 부적합판정 비율을 높이고 우량 수입식품에 대해서는 신속한 수입통관절차를 적용하여 수입식품통관 국가업무의 신속성과 효율성을 높이고자 한다. 다양한 예측모형의 민감도(sensitivity), 특이도(specificity), 수신자조작특성곡선(ROC, Receiver Operating Characteristic Curve) 등을 산출하여 AUROC(Area Under ROC Curve)를 예측모형의 성과 평가기준으로 Gaussian Naïve Bayes 예측모형을 제시하였다.

수입신고 현황 건수 대비 약 5% 정도를 무작위로 표본을 추출하여 정밀검사를 수행한다. 현재는 최근 3년동안의 수입현황을 수입국가별 부적합률, 수입건수, 부적합 건수, 부적합률, 위해물질 등의 검출여부를 반영하여 차등화된 위험점수를 산출하고 있다. 최근 5년동안 수입건수의 증가에 따라 정밀검사 건수 또한 증가하고 있지만 매년 부적합율은 좀처럼 개선되지 않고 있는 실정이다. 본 연구결과의 수입식품 사전예측모형을 적용하여 정밀검사 비중을 줄이거나 부적합율을 높여 수입통관 사무의 효율성과 경제성을 높일 수 있을 것이다.

본 연구에서는 수입식품의 불균형데이터를 정제, 분석하여 예측모형에 적합한 파생변수를 생성하였고 다양한 실험을 수행하였으며 제안한 모델의 예측성과를 제시하였다. 본 연구의 한계와 향후 연구방향을 정리하면 다음과 같다. 첫째, 정밀검사 결과에 대한 사전예측모형으로 Gaussian Naïve Bayes모형을 제안하였으나 향후 다양한 분류모형에 관한 연구가 필요할 것이다. 둘째, 수입식품데이터의 입력변수로 수입업소의 과거 행정처분이력, GMO식품여부, 수입통관 및 유통단계에서의 유해물질 등 검출여부 등을 추가로 고려하여 예측모형의 성과를 제고하여야 한다. 셋째, 수입식품데이터이외 전세계적으로 이슈가 되거나 전문뉴스 등에 등장하는 식품안전관련 사건사고 등의 정보를 수집 분석하여 실시간으로 수입검사관리에 반영하여 사전예측을 하여야 한다.

끝으로 본 논문에서 제안한 모형을 수입식품뿐만 아니라 국내식품현황자료에도 적용하여 새로운 예측모형의 적용 및 제안모형의 유용성에 대한 연구가 필요하다.

참 고 문 헌

[1] Allen, Bradley P., "Case-Based Reasoning: Business Applications", *Communications of the ACM,* Vol.37, No.3(1994), 40~42.

[2] Barandela, R., V. García, E. Rangel, and J. S. Sánchez, "Strategies for Learning in Class Imbalance Problems", *Pattern Recognition*, Vol.36, No.3(2003), 849~865.

[3] Beaver, William H., "Financial Ratios as Predictors of Failure", *Journal of Accounting Research*, Vol.4, No.3(1996), 71~111.

[4] Cao, Fuyuan, Jiye Liang, Deyu Li, and Xingwang Zhao, "A Weighting K-Modes Algorithm for Subspace Clustering of Categorical Data", *Neurocomputing*. Vol.108(2013) 23~30.

[5] Chang, D.S. and S.H. Lee, "A Study on the Us's Safety Control System for the Imported Food: Focused on the Processed Food", *The Journal of International Commerce,* Vol.31, No.4(2016), 325~50.

[6] Derrac, Joaquín, Chris Cornelis, Salvador García, and Francisco Herrera, "Enhancing Evolutionary Instance Selection Algorithms by Means of Fuzzy Rough Set Based Feature Selection", *Information Sciences*. Vol.186, No.1(2012), 73~92.

[7] Fawcett, Tom, "An Introduction to Roc Analysis", *PATTERN RECOGNITION LETTERS*, Vol.27, No.8(2006), 861~74.

[8] Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol.42, No.4(2012), 463~84.

[9] Ganganwar, Vaishali, "An Overview of Classification Algorithms for Imbalanced Datasets", *International Journal of Emerging Technology and Advanced Engineering*, Vol.2, No.4(2012), 42~47.

[10] García, V., A. I. Marqués, and J. S. Sánchez, "On the Use of Data Filtering Techniques for Credit Risk Prediction with Instance-Based Models", *Expert Systems With Applications*, Vol.39, No.18(2012), 267~76.

[11] Jin, Huang, and C. X. Ling, "Using Auc and Accuracy in Evaluating Learning Algorithms", IEEE Transactions on Knowledge and Data Engineering, Knowledge and Data Engineering, *IEEE Transactions on, IEEE Trans. Knowl. Data Eng*,Vol. 17, No.3(2005), 299~310.

[12] Jo, Taeho, and Nathalie Japkowicz, "Class Imbalances Versus Small Disjuncts", *SIGKDD Explor. Newsl*, Vol. 6, No.1(2004), 40~49.

[13] Kang, P.S., H.J. Lee and S.Z. Cho, " Svm Ensemble Techniques for Class Imbalance Problem", *KOREA INFORMATION SCIENCE SOCIETY*, Vol.31, No.2(2004), 706~708.

[14] Kim, U.M. and T.H. Hong, "The Prediction of Customers based on Case Based Reasoning with Weighted Factors for imbalanced Data Sets", *The Journal of Information Systems*, Vo1.21, No.1(2014), 29~45.

[15] Koufakou, Anna, and Michael Georgiopoulos, "A Fast Outlier Detection Strategy for Distributed High-Dimensional Data Sets with Mixed Attributes", *DATA MINING AND KNOWLEDGE DISCOVERY*, Vol.20, No.2(2010) 259~89.

[16] Lee, J.S. and J.G. Kwon, "A Hybrid Svm Classifier for Imbalanced Data Sets", Journal of Intelligence and Information Systems. Vol.19, No.2(2013), 125~40.

[17] Mac Namee, B., P. Cunningham, S. Byrne, and

O. I. Corrigan, "The Problem of Bias in Training Data in Regression Problems in Medical Decision Support", *ARTIFICIAL INTELLIGENCE IN MEDICINE,* Vol.24, No.1(2002), 51~70.

[18] Min, S.H., "Bankruptcy prediction using the improved bagging ensemble algorithm", Journal of Intelligence and Information Systems. Vol.20, No.4(2014), 121~139.

[19] Ohlson, James A., "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research,* Vol.18, No.1(1980), 109~31.

[20] Otey, M. E., A. Ghoting, and S. Parthasarathy, "Fast Distributed Outlier Detection in Mixed-Attribute Data Sets", *DATA MINING AND KNOWLEDGE DISCOVERY,* Vol.12, No.2-3 (2006), 203~28.

[21] Pai, Hao-Ting, Fan Wu, and Pei-Yun S. Hsueh, "A Relative Patterns Discovery for Enhancing Outlier Detection in Categorical Data", *DECISION SUPPORT SYSTEMS*, Vol.67(2014), 90~99.

[22] Park, M. S., H. N. Kim, and G. J. Bahk, "The Analysis of Food Safety Incidents in South Korea, 1998–2016", *Food Control*, Vol.81(2017), 196~199.

[23] Said, A. M., D. D. Dominic, and B. B. Samir, "Frequent Pattern-Based Outlier Detection Measurements: A Survey", *International Conference on Research and Innovation in Information Systems (ICRIIS)*, 2011

[24] Thompson, Christopher Glen, Rae Seon Kim, Ariel M. Aloe, and Betsy Jane Becker, "Extracting the Variance in Flation Factor and Other Multicollinearity Diagnostics from Typical Regression Results", *Basic & Applied Social Psychology*, Vol.39, No.2(2017), 81~90.

[25] Wu, X., V. Kumar, M. Steinbach, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, D. J. Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining", *Knowledge and Information Systems,* Vol.14, No.1(2008), 1~37.

[26] Zhao, Xingwang, Jiye Liang, and Fuyuan Cao, "A Simple and Effective Outlier Detection Algorithm for Categorical Data", *INTERNATIONAL JOURNAL OF MACHINE LEARNING AND CYBERNETICS*, Vol.5, No.3(2014), 469~77.

저 자 소 개

조 상 구(Cho, Sanggoo)
고려대학교 경영학사와 KAIST에서 MS를 취득하였다. 현재 식품안전정보원에서 주요 연구 관심분야는 응용통계, 성과관리 등이며, 현재 수입식품 안전관리 제도, 국가식품 안전관리 성과지표에 관한 연구를 진행 중이다.

최 경 현(Choi, Gyunghyun)
서강대학교 수학학사와 석사학위를 취득하였다. 미국 버지니아 공대 산업시스템공학과에서 M.S. 및 Ph.D. 학위를 취득하였다. 삼성SDS에서 근무하였으며, 현재 한양대학교 기술경영전문대학원에 재직 중이다. 주요 연구 관심분야는 기술전략 및 기획, 혁신경영, 제조 혁신 등이다.