ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 March: 21(1): 161~177 http://dx.doi.org/10.13088/jiis.2015.21.1.161

데이터마이닝을 이용한 허위거래 예측 모형: 농산물 도매시장 사례*

이선아
주식회사 웨이버스 기술본부
(salee@wavus.co.kr)

장남식
서울시립대학교 경영대학
(nchang@uos.ac.kr)

정보기술의 빠른 진화, 빅데이터의 등장, 분석기법의 고도화 등으로 인해 다량의 데이터로부터 의미있는 정보를 추출하는 데이터마이닝을 다양한 영역에 활용하고자 하는 시도들이 활발히 진행되고 있다. 그 중의 한 분야가 농산물 유통 영역인데, 농산물에 대한 지속적인 수요 증가와 전자경매의 활성화 등으로 수도권 농산물 도매시장에서만도 연간 수천만 건 이상의 거래가 이루어 진다. 그러나 급속한 거래량 증가와 더불어 과거로부터 관행적으로 이루어지고 있는 부정거래도 함께 증가하고 있는데 거래참가자들 사이의 결탁에 의해 발생하는 농산물 도매시장의 부정거래는 점차 지능화되는 추세이며, 이들을 감지하고 적발하기가 매우 어려운 실정이다. 이로 인해 농산물 유통환경의 공정거래 질서는 침해되고 시장에 대한 신뢰는 훼손되곤 한다. 따라서 거래투명성을 제고하고 유통비리를 구조적으로 개선하기 위한 과학적이고 자동화된 부정탐지시스템의 필요성이 어느 때보다도 절실히 요구되는 상황이다.

본 연구에서는 데이터마이닝의 의사결정나무를 이용하여 실제 발생하지 않은 거래를 실물 없이 거래한 것처럼 조작하여 대금을 정산하는 행위인 허위거래를 탐지하는 모형을 제시하였다. 이를 위해 실제 농산물 도매시장의 데이터를 수집하였고, 데이터의 정제 및 표준화 등의 선행작업을 수행하였다. 또한 변수 간의 상관관계 및 분포도 분석 등을 통해 데이터의 특성을 파악한 후 예측모형을 구축하여 허위거래와 정상거래를 분류하는 패턴을 도출하였으며, 최종적으로 시험용 데이터를 이용하여 모형을 평가하는 단계를 거쳐 결과의 적합성을 확인하였다. 향후 데이터마이닝을 이용한 부정탐지 모형을 허위거래뿐만 아니라 낙찰부정, 경매조작 등과 같이 다양화되는 부정거래에 적용하게 되면 보다 지대한 효과를 거둘 수 있으리라 사료된다.

주제어 : 농산물 도매시장, 부정탐지, 허위거래, 데이터마이닝, 의사결정나무

논문접수일 : 2014년 11월 25일 논문수정일 : 2015년 1월 6일 게재확정일 : 2015년 1월 14일
투고유형 : 국문일반 교신저자 : 장남식

1. 서론

정보통신 기술의 발전과 인터넷 모바일 기기의 이용 확산으로 인해 빅데이터에 대한 사회적 관심과 활용의 필요성이 크게 주목받고 있다. 3V(Volume, Variety, Velocity)로 정의되는 빅데이터는 무엇보다도 멀티미디어 콘텐츠, SNS를 통한 메시지 등과 같이 비정형 데이터의 종류가 다양화되고 양이 기하급수적으로 증가하고 있는 사회적·기술적 환경에 부합하여 탄생한 개념이라 해도 과언이 아니다(McKinsey, 2011; Stubbs, 2014). 많은 기업이나 기관들에서는 불확실한 경영 및 경제환경 하에서 전략적 우위를 선점하기 위해 빅데이터를 수집하고 분석하여 효과적으로

* 이 논문은 2013년도 서울시립대학교 학술연구조성비 지원에 의하여 연구되었음.

가치를 창출하고자 한다. 그러나 빅데이터의 개념이 소개되기 이전에도 다양한 산업 군에서는 이미 엄청난 양의 정형 데이터가 쌓여 왔으나 이에 대한 충분한 분석을 통한 활용 조차도 초보단계 수준에 불과한 것이 현실이다. 그리고 그 중의 하나가 바로 년간 천만 건 이상 발생하는 농산물 도매거래 영역이라 할 수 있다.

농산물에 대한 지속적인 수요 증가, 품목의 다각화, 경매의 전자화 등으로 인해 수도권 농산물 도매시장에서만도 하루에도 엄청난 양의 거래가 이루어 지고 있다. 그러나 급속한 거래량 증가와 더불어 과거로부터 관행적으로 이루어지고 있는 부정거래도 함께 증가하고 있는데, 거래참가자들 사이의 결탁에 의해 발생하게 되는 농산물 유통시장의 부정거래는 적발하기가 매우 어렵다. 물론 일부 부정거래는 피해자의 신고 또는 거래기록의 수작업 대조를 통해 감지할 수 있다. 하지만 부정거래의 유형이 점차 지능화되고 있는 상황에서 엄청난 양의 거래기록들을 담당부서나 담당자가 일일이 건 별로 조사하고 부정을 적발하는 것은 사실상 불가능하다. 따라서 거래투명성을 제고하고 유통비리를 구조적으로 개선하기 위한 과학적이고 자동화된 부정탐지시스템의 필요성 및 관련 연구가 어느 때보다도 절실한 상황이다.

지금까지 농산물 부정거래나 유통비리에 관련된 연구는 매우 미흡한 수준에 불과하다. 다만, 유통거래의 투명성 및 공정성에 관련된 연구가 일부 진행되었는데, 특히 전자식 경매가 가격결정 방식의 투명성 확보를 통해 경매담합 등의 경매비리에 대한 감시비용을 크게 감소시키는 중요한 역할을 하고 있다는 것을 강조한 연구가 있었다(Seo and Yang, 2011). 또한 농산물 도매시장의 당면과제와 이에 대한 개선방향의 일부로서 농산물 가격의 유동성 확대가 도매시장의 공정성 및 투명성 제고에 유효하리라는 연구 결과도 발표되었다(Wi and Kwon, 2009).

이에 반해 농산물 유통개선에 관련된 연구는 유통비리 관련 연구에 비해 상대적으로 활발히 수행되었는데, 정춘성(Jeong, 2000)은 농산물 중에서도 채소류 상품의 유통과정 상의 문제점을 파악하고 유통시장의 효율성 개선 방안을 제시하였다. 도매법인과 중도매인을 비롯한 매매참가인의 기능구분의 모호성으로 발생되는 문제와 가격결정 방법으로써의 경매・입찰제도의 실효성 문제에 대한 해결 방안으로 정가・수의매매의 효과적인 운용방안에 관한 연구도 있었다(Wi and Kwon, 2006). 이 외에도 유통구조나 정책의 문제점과 개선에 관한 다양한 연구 결과물들이 발표되었다(Choi et al., 2011; Sha, 2011).

한편, 대량의 데이터로부터 새롭고 의미있는 정보를 추출하는 데이터마이닝을 부정이나 사기 예측 모형 개발에 적용한 연구는 다수 진행되었는데, 차경엽(Cha, 2010)은 국민연금 부정수급 예측을 위해 의사결정나무 모형, 로지스틱 회귀모형, 인공신경망 모형 등을 개발하고 손해배상금 불성실 신고 도메인에 적용하여 의사결정나무 모형이 가장 우수한 예측력을 제공하였다는 결론을 도출하였다. 그리고 신용카드 부정사용 유형 중 하나인 현금불법융통 문제에 대해 데이터마이닝의 앙상블 모형의 유용성을 검증한 연구도 있었다(Song et al., 2007). 산업재해 발생 시 심사 과정과 급여지급 후에 부정수급으로 판명된 산재 청구 건을 데이터마이닝을 통해서 분석하여 부정수급의 유형을 발견한 연구도 시도되었는데. 이 연구에서는 총 61,536명의 최초요양 신청을 한 산재근로자 자료를 대상으로 다양한 데이터마이닝 기법을 오분류 비용 측면에서 비

교・평가하였다(Ham and Hong, 2008). 김태형과 김영화(Kim and Kim, 2013)는 여신기관을 대상으로 신용 및 카드 대출심사에 데이터마이닝을 적용하여 총 9가지의 분류모형을 구축하고 비교하였다. 이 밖에도 은행부실이나 기업도산 예측 등에 다수의 관련 연구가 진행되었다(Sung et al., 1999; Tam and Kiang, 1992).

본 연구에서는 데이터마이닝의 의사결정나무를 이용하여 실제 발생하지 않은 거래를 실물 없이 거래한 것처럼 조작하여 대금을 정산하는 행위인 허위거래를 탐지하는 모형을 제시하였다. 이를 위해 실제 농산물 도매시장의 데이터를 수집하여 데이터의 특성을 분석하였고, 모형을 통해 허위거래와 정상거래를 분류하는 패턴을 도출한 후, 시험용 데이터를 이용하여 모형을 평가하는 단계를 거쳐 결과의 적합성을 확인하였다(Lee, 2013).

논문의 구성은 다음과 같다. 2장에서는 농산물 도매시장의 거래형태, 유통 종사자의 역할, 부정거래의 유형 및 특성 등 농산물 유통시장 환경에 대해 조사하였다. 3장에서는 실증분석을 위해 수집한 농산물 거래데이터의 특성을 기초통계분석을 통해 살펴보았고, 데이터 정제 및 파생변수 생성, 실증분석의 범위 및 연구방법 등을 논하였다. 4장에서는 위에서 준비된 데이터를 대상으로 의사결정나무 기법을 이용하여 허위거래 예측모형을 구축하고 평가하였으며, 5장에서는 연구의 결과와 시사점을 정리하였다.

2. 농산물 유통시장 환경

2.1 유통 종사자 및 거래 형태

농산물 도매시장의 거래방식은 '거래 총수 최소화의 원리'를 기반으로 한다. 이것은 농산물 거래를 생산자와 소비자가 직접 거래할 때의 총수보다 생산자와 소비자 사이에 도매시장이 개입하여 거래를 할 경우 거래하는 총수가 줄어드는 원리로 거래비용을 절감하는 효과를 기대할 수 있다. 또한 생산자는 생산에만, 상인은 유통에만 전념함으로써 분업에 의한 운영의 효율성 제고가 가능하다(Garak, 2014). 도매시장 유통 종사자는 크게 도매시장법인, 중도매인, 매매참가인, 산지유통인, 출하자, 경매사 등으로 구성되는데, 다음과 같은 역할을 수행하게 된다(Egmarket, 2014).

- 도매시장법인 : 농산물을 위탁 받아 상장하여 중도매인이나 매매참가인에게 판매를 대행하는 자 또는 조직. 위탁 받은 농산물을 높은 가격에 판매할수록 수수료 수입이 증가하여, 결과적으로 출하자의 이익을 대변하는 역할을 수행함. 또한 거래 전반에 걸쳐 발생하는 데이터를 저장하고 관리함
- 중도매인 : 도매시장 또는 공판장 개설자의 허가 또는 지정을 받아 도매시장 또는 공판장에서 상장된 농산물을 매수하여 도매거래하거나 매매를 중개하는 자
- 매매참가인 : 도매시장법인에 등록하고 도매시장 또는 공판장에 상장된 농산물을 직접 매수하는 가공업자, 농산물 소매업자, 소비자 단체 등의 수요자
- 산지유통인 : 도매시장법인 또는 공판장 개

설자에게 등록하고 농산물을 수집하여 도매시장 또는 공판장에 출하하는 영업을 하는 자
- 출하자 : 도매시장에 농산물을 출하하는 생산자나 생산자 단체
- 경매사 : 도매시장법인에 의해 상장된 농산물에 대한 경매 우선 순위를 정하고, 상장 농산물의 가격평가 및 경락자 결정을 수행하는 자

농산물 도매시장의 거래 형태는 '상장거래'와 '상장예외거래'로 구분되며, '상장거래'는 다시 '경매'와 '정가・수의매매'로 나뉜다. 통계에 따르면 도매시장 거래 중 '상장거래'의 비중이 약 90%인 반면 '상장예외거래'는 10% 내외에 불과하다. 또한 '상장거래' 중에서도 '경매'의 비중이 약 92%를 차지하는데, 본 연구의 부정거래 예측 모형의 대상이 되는 '정가・수의매매'의 특징을 살펴보면 다음과 같다(Kim et al., 2009).

- 전제조건 : 공급이 탄력적이며 수요를 초과하는 품목으로, 주로 대형소매점이나 식품기업을 대상으로 매매가 이루어 짐
- 대상품목 : 시설 채소류와 저장성이 있는 과일류
- 장・단점 : 거래의 안정성에 대한 대응이 가능하고 거래에 시간적인 제약을 받지 않는 반면, 투명한 거래가 되지 못할 가능성이 높음

2.2 유통시장의 부정거래

농산물 유통시장에서의 부정은 농산물 거래 참가자들이 결탁하여 적법하지 않은 행위를 통해 이익을 내는 것을 의미한다. 일반적으로 부정은 내부통제시스템이 취약한 환경에서 각종 편법을 이용하여 경제적 이익을 취하고자 하는 심리가 작용하여 발생하는 경우가 많다. 그 밖에 최저 거래실적이나 경영목표 달성에 대한 압박으로 인한 부담감도 부정의 원인이 되곤 한다.

농산물 도매거래에서 발생할 수 있는 부정거래들은 거래 프로세스 내에서 데이터 확인만을 통해 부정을 탐지할 수 있는 유형과 거래 데이터만으로는 부정을 탐지할 수 없는 유형으로 분류할 수 있는데 '최고가격 제시자 외 낙찰', '일괄경매', '편중낙찰', '경매가 조작', '편법 저가낙찰' 등이 전자에, '정산가 임의정정', '허위거래', '기록상장' 등이 후자에 해당된다. 각 부정거래에 대한 정의는 다음과 같다.

- 최고가격 제시자 외 낙찰 : 전자경매 과정에서 최고가격을 제시한 중도매인 이외의 중도매인에게 낙찰이 이루어지는 행위
- 일괄경매 : 동일 생산자 출하물량 중 과수별 개별경매를 하지 않고, 중심이 되는 과수만 경매한 뒤 나머지 등급은 경매를 생략한 채 기리 폭을 임의로 정하여 가격을 결정하는 행위
- 편중낙찰 : 전자경매 과정에서 특정 출하자의 상품을 특정 중도매인이 집중적으로 낙찰 받는 행위
- 경매가 조작 : 경매 결과 데이터 중 경락가를 임의로 조작하는 행위
- 편법 저가낙찰 : 특정 중도매인이 1차 경매에서 일반적인 가격보다 높게 낙찰 받은 후, 재경매를 요구하여 2차 경매에서 낮은 가격으로 낙찰 받는 행위
- 정산가 임의정정 : 상장예외품목을 취급하는 중도매인이 판매한 금액보다 낮은 가격으로 출하자에게 대금을 정산하는 행위

- 허위거래 : 실제 발생하지 않는 거래를 실물 없이 거래한 것처럼 속여 대금을 정산하는 행위
- 기록상장 : 실물은 존재하나 경매가 아닌 방법으로 거래한 것을 경매로 한 것처럼 꾸미는 행위

예를 들어, '최고가격 제시자 외 낙찰'은 데이터의 확인을 통해 최고 가격을 제시하지 않은 중도매인이 낙찰 받은 거래를 추출하여 탐지가 가능하다. 반면에 '허위거래'의 경우는 이미 부정이 발생한 이후 정산자료가 입력되기 때문에 데이터 내에서는 부정거래를 탐지할 수가 없다. 본 연구에서는 데이터마이닝의 분류모형(classification)의 일환으로 거래부정 예측모형을 구축하고자 하였다. 이를 위해서는 데이터에 '부정' 또는 '정상' 거래를 구분하여 주는 목표변수가 반드시 필요하나 '정산가 임의정정'과 '기록상장'의 경우 목표변수의 수집이 가능하지 않았기 때문에 목표변수의 수집이 가능했던 '허위거래'만을 대상으로 실증분석을 수행하였다.

3. 데이터 수집 및 분석범위

3.1 데이터 수집 및 구성

6개 농산물 도매시장법인으로부터 20xx년 1월부터 32개월 간의 청과거래 정산데이터 42,313,819건을 수집하여 데이터마트를 구축하였다. 이 데이터는 총 34개의 변수로 구성되어 있었으나 동일한 값이 기록되어 있는 변수, 빈 값을 다수 포함하고 있는 변수, 개인정보를 포함하고 있는 변수, 단순히 거래레코드를 식별하기 위한 변수 등을 제외한 후, <Table 1>과 같이 총

〈Table 1〉 Variable List

Name	Description
Auction Place (경매장구분)	Auction Place Code
Auction Type (경매구분)	Auction Type Code
Auction Time (경락시간)	Recoded Auction Time(Year-Month-Date-Time)
Item Code (품목코드)	Item Code
Item Name (품목명)	Item Name
Number of Unit (거래단량)	Number or Volume per Unit Package
Unit Code (단위코드)	Standard Unit Code for Item
Number of Package (거래수량)	Trade Amount measured in Number of Package
Auction Price (경락가)	Contract Price per Package
Producer Name (출하처명)	Name of Item Producer
Sales Type (판매구분)	Sales Type Code
Listed Status (상장구분)	Listed or Unlisted
Trade Volume (거래물량)	Number of Unit * Number of Package
Trade Amount (거래금액)	Auction Price * Number of Package
Unloading Cost (하역비)	Unloading Cost
Commission (위탁수수료)	Commission for Selling Products

〈Table 2〉 Derived Variable List

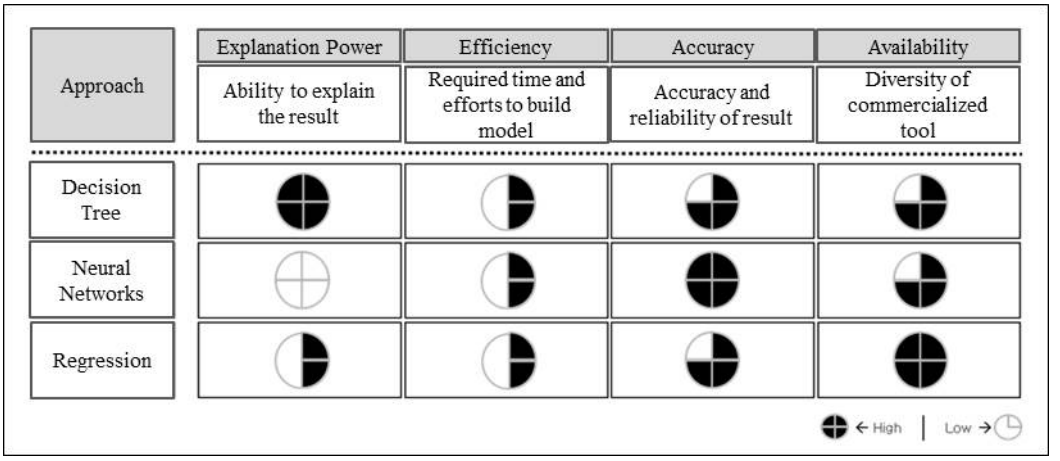
Name	Definition
ATA(Average Trade Amount)	Average of trade amount per month for each item
ATV(Average Trade Volume)	Average of trade volume per month for each item
RTA(Ratio of Trade Amount)	(Standard deviation of trade amount/ATA) per month for each item
AUTA(Average Unit Trade Amount)	Average of (trade amount/trade volume) per month for each item
RUTA(Ratio of Unit Trade Amount)	(Standard deviation of unit trade amount)/(AUTA) per month for each item
ATAP(Average Trade Amount per Producer)	ATA per producer
ATVP(Average Trade Volume per Producer)	ATV per producer
RTAP(Ratio of Trade Amount per Producer)	(Standard deviation of trade amount per producer/ATAP) per month for each item
AUTAP(Average Unit Trade Amount per Producer)	AUTA per producer
RUTAP(Ratio of Unit Trade Amount per Producer)	(Standard deviation of unit trade amount per producer/AUTAP) per month for each item
ATAW(Average Trade Amount per intermediary Wholesaler)	ATA per intermediary wholesaler
ATVW(Average Trade Volume per intermediary Wholesaler)	ATV per intermediary wholesaler
RTAW(Ratio of Trade Amount per intermediary Wholesaler)	(Standard deviation of trade amount per intermediary wholesaler/ATAW) per month for each item
AUTAW(Average Unit Trade Amount per intermediary Wholesaler)	AUTA per intermediary wholesaler
RUTAW(Ratio of Unit Trade Amount per intermediary Wholesaler)	(Standard deviation of unit trade amount per intermediary wholesaler /AUTAW) per month for each item

16개의 변수를 선정하였다.

3.2 데이터 표준화

일반적으로 서로 다른 원천에서 수집한 데이터를 통합하여 분석할 경우 동일 변수 값 간에 다양한 의미의 충돌이 발생하게 되며, 이는 곧 분석 결과의 품질을 저하시키는 원인이 되곤 한다. 박진수(Park, 2006)는 데이터 레벨에 의한 의미충돌을 '데이터 값', '데이터 표현', '데이터 단위', '데이터 정밀도', '데이터 값 신뢰성', '공간 도메인' 등으로 분류하였다.

본 연구를 위해 수집한 데이터도 6개의 도매시장법인이 독립적으로 시스템을 운영하여 기록함에 따라 데이터 레벨에 의한 의미충돌이 발견되었으며, 주로 유사한 객체가 다른 데이터 타입 혹은 데이터 포맷 형식으로 표현됨으로써 발생하는 '데이터 표현' 충돌이 다수였다. 문제 해결을 위해 먼저 시간 표기 방식을 'YYYYMMDDHHMMSS'라는 하나의 형태로 표준화하였다. 다음으로 변수들 중에서 '품목명'의 경우 도매시장법인 별로 동일한 품목을 다른 이름으로 입력·저장하고 있어, 이에 대한 표준화 작업을 수행하였다. 예를 들어 오렌지의 경우, '오렌지', '오렌지.', '오렌지(수입)', '오렌지(오렌지)' 등과 같이 다양한



〈Figure 1〉 Comparison of Data Mining Approaches

형태로 기록된 품목명을 '오렌지'라는 동일 품목명으로 통일하였다.

3.3 파생변수 생성

기존의 변수를 이용하여 거래부정에 유의하다고 판단되는 파생변수 15개를 <Table 2>와 같이 추가적으로 생성하였다. 예를 들어 '1단위 거래금액 평균(AUTA: Average Unit Trade Amount)' 변수는 기존에 존재하는 거래금액을 거래물량으로 나누어 거래한 물량 1단위 당의 거래금액을 산출한 것이고, '(월 품목) 거래금액 평균(ATA: Average Trade Amount)' 변수는 월별, 품목별 거래금액의 평균이다.

3.4 분석 범위 및 방법

실증분석에서는 수입자유화로 인해 최근 거래물량이 급증하고 있는 '수입과일' 품목을 대상으로 '판매원표 허위작성' 거래에 대한 분석을 수행하였다. 또한 경매구분 값이 '전자경매', '비상장거래', '정가·수의매매' 중에서 '정가·수의매매'이고, 품목코드가 수입과일[1]*을 의미하는 품목들을 선정하였으며, 거래일은 최근 1년으로 한정하였다. 최종적으로 25,171건의 레코드가 추출되었으며, 171건을 '허위거래'로 나머지 25,000건을 '정상거래'로 분류하였다. 특히 '허위거래'로 분류한 171건은 실제로 판명된 허위거래와 현업전문가들의 의견을 반영하여 허위로 의심되는 2개의 출하처에서 발생한 거래들이다.

허위거래 예측을 위해 실증분석에 필요한 데이터마이닝 작업유형은 부류 값(허위, 정상)이 포함된 과거의 데이터로부터 부류 별 특성을 찾아내어 모형을 만들고, 이를 토대로 새로운 거래의 부류 값을 예측하는 분류작업이다. 일반적으로 분류작업에 사용하는 데이터마이닝 기법으로는 '의사결정나무', '신경망', '회귀분석' 등이 있는데, 본 연구에서는 분류나 예측의 근거를 알 수 있고, 어떠한 속성(변수)들이 각각의 부류 값에 결정적인 영향을 주는가를 쉽게 파악하고자 하는 현장의 요구를 수렴하여 '의사결정나무'를

* 오렌지, 체리, 포도, 파인애플, 키위, 망고, 레몬, 석류, 자몽, 아보카도 등을 포함한다.

데이터마이닝 기법으로 선정하였다. <Figure 1>은 데이터마이닝 기법들을 '모형의 설명력', '작업의 효율성', '모형의 정확성', '기법의 유용성' 측면에서 비교·평가한 것이다(Chang et al., 1999).

4. 실증분석

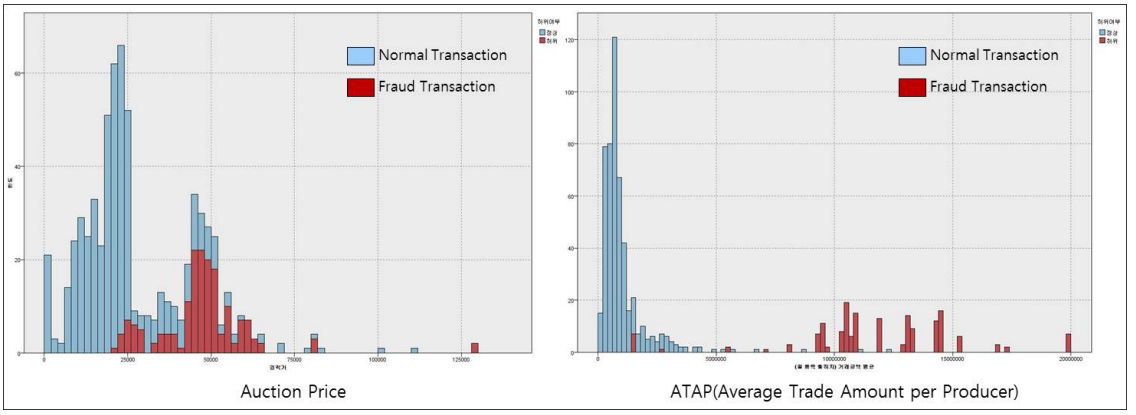
실증분석에 사용된 변수는 파생변수를 포함한 31개 중 20개와 목표변수 등으로 총 21개이다. <Table 2>에서 선정했던 최초 입력변수 중에서 '경매구분', '품목코드'와 같이 분석의 범위를 정의하는데 사용했던 변수들과 '거래수량', '거래단량'과 같이 파생변수를 생성하는데 사용했던 최초 입력변수 11개는 더 이상 자체적인 의미가 없다고 판단하여 제외하였다. 실증분석은 '기초통계분석' à '예측모델링' à '모형평가' 순으로 진행하였으며, SPSS Modeler를 사용하였다.

4.1 기초 통계분석

기초 통계분석은 데이터에 대한 이해도를 높이고, 허위 및 정상 거래에 영향을 미치는 변수들의 변별력을 파악하고자 하는 목적으로, 개별 변수들의 평균 및 표준편차를 비교한 후 변수 간의 상관관계 분석하는 단계로 수행하였다. 이 단계를 통해 파악된 변수들은 향후 허위거래 예측모델링에 주요한 역할을 할 것으로 예견하였으며, 실제로 예측모델의 신뢰도 및 설명력 제고에

〈Table 3〉 Comparison of Average of Variables

	(A) Fraud (171)	(B) Normal (25,000)	A/B (%)
Auction Price	46,689	22,951	203
Trade Volume	2,668	532	502
Trade Amount	11,984,056	1,015,626	1,180
Unloading Cost	48,251	9,493	508
Commission	479,362	40,887	1,172
ATA(Average Trade Amount)	2,097,313	1,010,560	208
ATV(Average Trade Volume)	533	541	99
RTA(Ratio of Trade Amount)	605	98	616
AUTA(Average Unit Trade Amount)	5,022	2,175	231
RUTA(Ratio of Unit Trade Amount)	119	107	112
ATAP(Average Trade Amount per Producer)	11,984,056	1,003,756	1,194
ATVP(Average Trade Volume per Producer)	2,668	530	504
RTAP(Ratio of Trade Amount per Producer)	100	99	101
AUTAP(Average Unit Trade Amount per Producer)	5,866	2,221	264
RUTAP(Ratio of Unit Trade Amount per Producer)	100	104	97
ATAW(Average Trade Amount per int. Wholesaler) Wholesaler)Wholesaler)	11,191,661	1,004,041	1,115
ATVW(Average Trade Volume per int. Wholesaler)	2,443	526	464
RTAW(Ratio of Trade Amount per int. Wholesaler)	115	101	115
AUTAW(Average Unit Trade Amount per int. Wholesaler) Wholesaler)intermediary Wholesaler)	5,866	2,227	263
RUTAW(Ratio of Unit Trade Amount per int. Wholesaler) Wholesaler)	100	104	96



〈Figure 2〉 Distribution of Major Variables

도움이 되었다.

<Table 3>은 허위거래와 정상거래에 대한 입력변수들의 평균 값을 비교한 것으로, 예를 들어 "허위거래와 정상거래의 경락가(Auction Price) 평균 값은 각각 46,689원과 22,951원으로 허위거래의 경락가 평균 값이 정상거래에 비해 203% 크다"라고 해석한다. 이 표에 따르면 '(월 품목) 거래물량 평균(ATV)'과 '(월 품목 출하자) 1단위 거래금액의 평균과의 편차비율(RTAP)'의 정상거래 대비 허위거래 평균 값의 비율은 각각 99%, 101%로 차이가 거의 없는 것으로 나타났다. 반면에 '거래금액(Trade Amount)', '위탁수수료(Commission)', '(월 품목 출하자) 거래금액 평균(ATAP)', '(월 품목 중도매인) 거래금액 평균(ATAW)'의 정상거래 대비 허위거래 평균 값 비율은 1,000%이상으로 이들 변수들이 허위거래와 정상거래를 구분하는 핵심 변수라는 것을 알 수 있다. 이것은 향후 의사결정나무를 이용하여 모델링을 수행할 때 핵심변수 중 일부가 의사결정나무의 뿌리마디를 구성하리라는 것을 짐작하게 한다.

평균 값의 차이는 <Figure 2>의 변수 값 분포도에서도 확연히 구분할 수 있는데 정상거래 대비 허위거래 평균 값의 비율이 203%인 '경락가(Auction Price)'의 경우 허위거래의 '경략가'가 주로 정상거래의 상위 값 범위 내에서 겹쳐서 분포하고 있다. 그러나 정상거래 대비 허위거래 평균 값의 비율이 1,194%인 '(월 품목 출하자) 거래금액 평균(ATAP)'은 허위거래의 값들이 정상거래의 최상위 값보다 우측에 넓게 분포함으로써 정상거래에 비해 평균 값과 표준편차가 확연히 크다는 것을 보여준다.

<Table 4>는 주요 입력변수간의 상관계수를 산출한 매트릭스로서 <Table 3>과 병행해서 해석하면 다음 장에서 구축하게 되는 예측모형에 대한 이론적 근거를 제시할 수 있다. 예를 들어, 정상거래 대비 허위거래 평균 값 비율이 1,194%로 가장 높았던 'F: (월 품목 출하자) 거래금액 평균(ATAP)'이 의사결정나무 모형의 뿌리마디로 선정되게 되면 이 변수와 상관계수가 상대적으로 높은 'A: 거래금액(Trade Amount)', 'D: (월 품목) 거래금액 평균과의 편차 비율(RTA)', 'G: (월 품목 출하자) 거래물량 평균(ATVP)', 'I: (월 품목 중도매인) 거래금액 평균(ATAW)' 등은 최

〈Table 4〉 Correlation Coefficient

	A	B	C	D	E	F	G	H	I	J
A	-	0.61	0.01	0.91	0.12	0.91	0.69	0.23	0.90	0.51
B	0.61	-	0.44	0.37	-0.01	0.65	0.44	0.06	0.64	0.49
C	0.01	0.44	-	-0.09	-0.07	0.00	0.26	0.01	-0.02	-0.40
D	0.91	0.37	-0.09	-	0.11	0.82	0.74	0.32	0.82	0.40
E	0.12	-0.01	-0.07	0.11	-	0.12	-0.01	-0.01	0.12	0.15
F	0.91	0.65	0.00	0.82	0.12	-	0.76	0.04	0.96	0.55
G	0.69	0.44	0.26	0.74	-0.01	0.76	-	0.02	0.70	0.09
H	0.23	0.06	0.01	0.32	-0.01	0.04	0.02	-	0.09	0.09
I	0.90	0.64	-0.02	0.82	0.12	0.96	0.70	0.09	-	0.56
J	0.51	0.49	-0.40	0.40	0.15	0.55	0.09	0.09	0.56	-

A: Trade Amount
B: ATA(Average Trade Amount)
C: ATV(Average Trade Volume)
D: RTA(Ratio of Trade Amount)
E: RUTA(Ratio of Unit Trade Amount)
F: ATAP(Average Trade Amount per Producer)
G: ATVP(Average Trade Volume per Producer)
H: RTAP(Ratio of Trade Amount per Producer)
I: ATAW(Average Trade Amount per intermediary Wholesaler)
J: AUTAW(Average Unit Trade Amount per intermediary Wholesaler)

종모형에 포함되지 않을 확률이 큰데, 이것은 변수간의 상호의존도가 높다는 '다중공선성(multicollinearity)'이 존재할 수 있다는 것을 의미한다(Rho, 1998). 즉, 상호의존도가 높은 변수들이 예측모형에 함께 포함될 경우 모형의 예측력 및 설명력을 저하시키는 문제가 발생하기 때문에 정상거래 대비 허위거래 평균 값 비율이 각각 1,180%, 1,115%로 큰 변별력을 지녔던 'A: 거래금액(Trade Amount)', 'I: (월 품목 중도매인) 거래금액 평균(ATAW)' 등은 'F: (월 품목 출하자) 거래금액 평균(ATAP)'과 함께 최종모형에 포함되지 않을 확률이 높다고 볼 수 있다.

4.2 예측 모델링

의사결정나무는 분류나 예측에 변별력이 낮은 변수들을 모형 구축 시 자체적으로 배제시키므로 기초 통계분석에서 살펴 본 20개의 모든 변수들을 입력변수로 정하였다. 그러나 거래 건수에 있어서는 허위거래의 수(171건)가 전체 데이터(25,171)에서 차지하는 비율이 0.68%에 불과해 허위거래의 패턴을 도출하기가 어렵다고 판단하여 허위거래의 수와 정상거래의 수의 비율을 각각 25%와 75%로 결정하였다. 따라서 최종 모델링에 사용한 데이터의 수는 허위거래 171건과 정상거래 500건으로 총 671건이며, 정상거래 25,000건 중 임의 표본 추출을 통해 500건의 정상거래를 선정하였다. 또한 예측모형의 구축과 모형의 예측력 시험을 위해 전체 데이터를 <Table 5>와 같이 모형 추정용(training) 데이터와 모형 시험용(test) 데이터로 구성하였다. 이와 더불어 임의 표본 추출과정에서 발생하는 데이

터의 치우침(bias)을 예방하고, 결과의 일반화를 도모하기 위해 위의 방법을 3회 반복하여 3개의 데이터 세트를 만들었다.

〈Table 5〉 Data Set

	Fraud	Normal	Total
Training Data Set	137	400	537
Test Data Set	34	100	134
Total	171	500	671

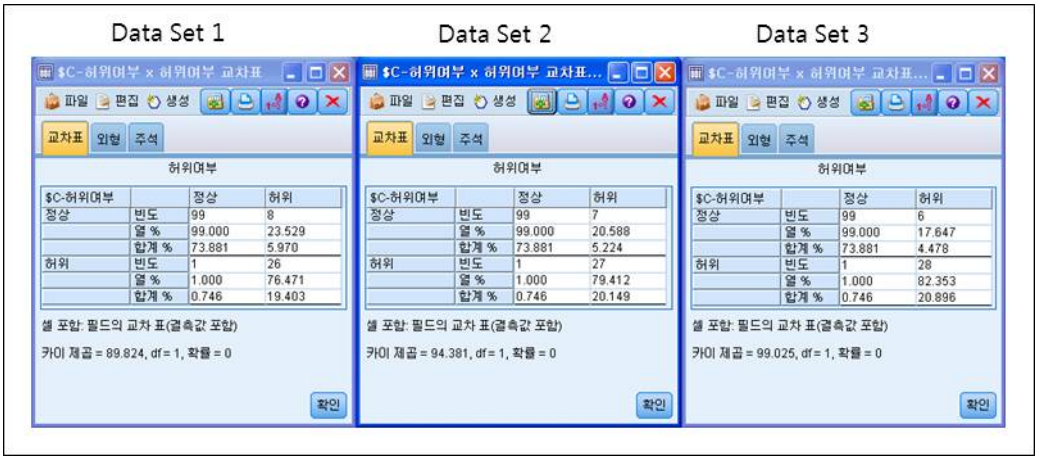
모형 구축 시 모형의 과잉맞춤(overfit)으로 인한 예측력 저하를 방지하기 위해 마디의 최저 순수도 및 최소 관측 개수를 이용한 가지치기(pruning) 방식을 택하였다. 이 값들은 모형 구축 전에 미리 정하는데, 마디의 최저 순수도는 마디를 구성하는 사례들에서 한 종류의 부류 값에 속한 사례의 비율이 사전에 정의한 최저 순수도보다 커지면 나무의 확장을 중지하는 방식이며, 최소 관측개수는 끝마디에 포함된 사례의 개수가 정의된 값 이하가 되면 확장을 중지하는 방식이다. 따라서 최저 순수도의 값을 낮게 정의할수록, 또한 최소 관측개수의 값을 크게 할수록 나무의 구조는 단순화된다(Chang, 2005). 본 연구에서는 위의 두 가지 중 먼저 만족되는 기준에 도달하면 나무의 확장을 멈추었으며, 최종적으로 <Table 6>과 같이 4개의 주요규칙(룰)을 도출하였다.

결과적으로 최종 모형에서는 기초 통계분석에서 예측된 바와 같이 '(월 품목 출하자) 거래금액 평균(ATAP)'이 허위거래와 정상거래를 구분하는 주요 변수로 선정되었다. 특히 <Table 3>의 변수 평균 값 비교표에서 정상거래 대비 허위거래 평균 값의 비율 각각 203%와 207%로 상대적으로 다른 변수들에 비해 비율이 크지 않았던 '경락가(Auction Price)'와 '(월 품목) 거래금액 평균(ATA)'이 '(월 품목 출하자) 거래금액 평균(ATAP)'과 조합하여 허위거래를 구분하는 주요 규칙을 구성하였다. 정상거래의 패턴에 있어서도 정상거래 대비 허위거래 평균 값의 비율 각각 99%와 96%로 변별력이 없다고 추정했던 '(월 품목) 거래물량 평균(ATV)'과 '(월 품목 중도매인) 1단위 거래금액 평균(AUTAW)'이 규칙에 포함되어 있음을 알 수 있다. <Table 6>의 규칙은 데이터 세트 1의 모형 구축용 데이터에 의해 도출된 것으로 데이터 세트 2 및 3에서도 동일한 변

〈Table 6〉 Rules Derived from Decision Tree

Transaction Type	Rules	Total*
Fraud (137)	ATAP(Average Trade Amount per Producer) > 5,343,000 & Auction Price > 35,000	133
	ATAP(Average Trade Amount per Producer) > 5,343,000 & Auction Price ≤ 35,000 & ATA(Average Trade Amount) ≤ 2,622,110	
Normal (400)	ATAP(Average Trade Amount per Producer) ≤ 5,343,000 & ATV(Average Trade Volume) ≤ 871	398
	ATAP(Average Trade Amount per Producer) ≤ 5,343,000 & ATV(Average Trade Volume) > 871 & AUTAW(Average Unit Trade Amount per intermediary Wholesaler) > 1,572	

* Total number of transactions applied to the rules



〈Figure 3〉 Prediction Accuracy

수들이 규칙에 포함되었다. 다만 규칙을 구성하는 값들 사이에는 약간의 차이가 있었는데, 이것은 위의 규칙들에서 단단위 및 십단위 값들이 큰 의미가 없음을 시사한다.

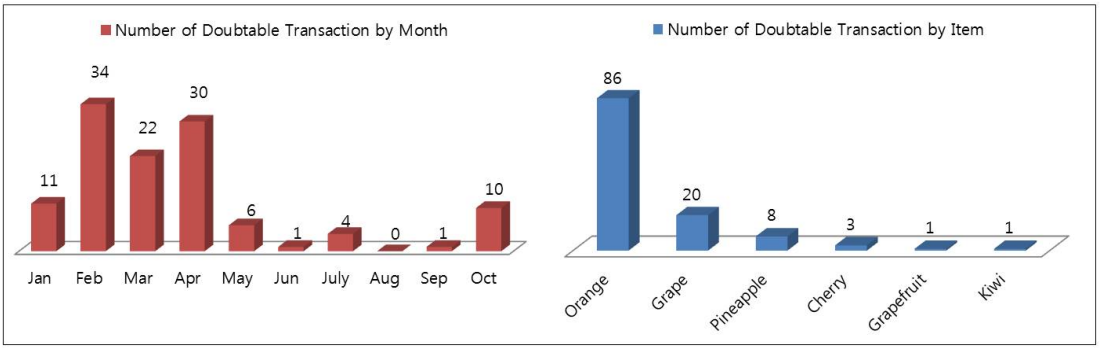
4.3 모형 평가

위에서 구축한 모형을 정분류율에 의해 평가하였다. 정분류율은 입력 변수의 값을 갖고 각 사례의 실제 결과 (범주)를 예측하는 분류적 예측모형의 성능평가에 가장 일반적으로 사용되어 온 방법으로 이분적 분류 능력에 대한 평가라고 할 수 있다(Chang, 2005; Tam and Kiang, 1992). <Figure 3>은 3개의 데이터 세트에서 구축한 각각의 모형을 허위거래 34건, 정상거래 100건으로 구성된 각각의 모형시험용 데이터에 적용한 결과이다. 예를 들어, 데이터 세트 2의 경우 100건의 정상거래 중 99건을 정상거래로, 34건의 허위거래 중 27건을 허위거래로 정분류한 것으로 나타났다. 결과적으로 3개의 데이터 세트에 대한 모형의 정상거래 정분류율은 평균 99%, 허위거래의 정분류율은 평균 79%로 나타났다.

5. 결론 및 시사점

본 연구에서는 데이터마이닝의 의사결정나무를 이용하여 실제 발생하지 않은 거래를 실물 없이 거래한 것처럼 조작하여 대금을 정산하는 행위인 허위거래를 탐지하는 모형을 제시하였다. 이를 위해 실제 농산물 도매시장의 데이터를 수집하여 표준화와 파생변수 생성 등의 데이터 정제 및 보강 작업을 선 수행하였으며, 기초 통계분석 à 예측모델링 à 모형평가의 단계를 거쳐 예측모형을 구축하고 결과의 적합성을 확인하였다.

월별로 10만건 이상 발생하는 거래를 담당자가 일일이 검토하여 허위거래를 적발하는 것은 비효율적일 뿐만 아니라 실제적으로 불가능한 작업이다. <Figure 4>는 최근 10개월동안 발생한 38,849건의 수입과일 거래에 데이터마이닝 모델을 적용한 것으로, 전체 거래의 0.31%에 해당하



〈Figure 4〉 Number of Doubtable Transaction

는 총 119건의 허위 의심거래를 추출하였고 이것을 월별・품목별로 정리한 것이다. 월별로는 2월과 4월에 각각 34건과 30건으로 가장 많은 허위의심거래가 탐지되었으며, 품목별로는 오렌지 거래가 86건, 포도 거래가 20건으로 오렌지와 포도가 전체 의심거래의 89%를 차지했다. 이 같은 결과로 토대로 본 연구의 기여도는 모든 거래를 확인하는 것보다는 선택과 집중 차원에서 허위의심거래 군을 추출하고, 이 거래 군에 대해 심도있는 검증 및 확인 작업을 수행하는 것이 보다 효율적인 방법이라는 것을 실증분석을 통해 제시하였다는 데에 있다. 향후 이러한 연구를 허위거래뿐만 아니라 낙찰부정, 경매조작 등과 같이 다양화되는 부정거래에 적용하게 되면 보다 지대한 효과를 거둘 수 있으리라 사료된다.

참고 문헌(References)

Cha, K. Y., "An Application of Data-Mining Tool in Fraud Pension Payment Prediction," *Communications for Statistical Applications and Methods*, Vol.17, No.1(2010), 1~8.

Chang, N., "Improving the Effect of Customer Classification Models: A Pre-segmentation Approach," *Information Systems Review*, Vol.7, No.2(2005), 23~40.

Chang, N., S. W. Hong, and J. H. Jang, *Data Mining*, Daecheong, 1999.

Choi, S. -H., J. -W. Kim, K. -R. Kim, and Y. S. Lee, "A Study on the Problem and Improvement of Farm Product Structure in Korea," *Journal of Franchise Management*, Vol.2, No.2(2011), 70~83.

Egmarket, *Distributor's Role*, Available at http://egmarket.busan.go.kr/02_currency/02_01.jsp (Accessed 20 September, 2014).

Garak, *Market Function*, Available at http://www.garak.co.kr/gongsa/jsp/mk/marketinfo/overview.jsp (Accessed 18 August, 2014).

Ham, S. O. and J. S. Hong, "A Study on the Fraud Detection of Industrial Accident Compensation Insurance," *Proceedings of 2008 KORMS Fall Conference*, (2008), 342~345.

Jeong, C. S., "A Study on the Agricultural Product Market: The Case of Vegetable Products," *Master's Thesis*, Department of Economics, Kyung Hee University, 2000.

Kim, D. W., J. W. Song, D. S. Kim, J. H. Park,

H. N. Park and Y. R. Lee, “Improving Sales Efforts of Intermediary Wholesaler in Garak Market,” *Research Report*, Seoul Agro-Fisheries & Food Corporation, 2009.

Kim, T. -H and Y. -H. Kim, “A Study on the Analysis of Customer Loan for the Credit Finance Company Using Classification Model,” *Journal of the Korean Data & Information Science Society*, Vol.24, No.3(2013), 411~425.

Lee, S. A., “A Study on the Fraud Detection using Data Mining: The Case of Agricultural Products Distribution Market,” *Master’s Thesis*, College of Business Administration, University of Seoul, 2013.

McKinsey Global Institute, “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” McKinsey and Company, 2011.

Park, J., “Real-time Data Integration using Ontology and Semantic Mediators,” *Asia Pacific Journal of Information Systems*, Vol. 16, No.4(2006), 151~178.

Rho, B. H., J. H. Min, and G. H. Lee, *Introduction to Statistics*, Bobmunsa, 1998.

Seo, K. N. and S. R. Yang, “The Effect of the Electronic Auction on the Price Efficiency in the Garak Market,” *Korean Journal of Agricultural Management and Policy*, Vol.38, No.2(2011), 175~195.

Sha, D. C., “The Legislation on the Stability of Supply and Reform of Circulation Structure on Agricultural Products,” *Hongik Law Review*, Vol.12, No.2(2011), 167~193.

Song, Y., W. Han and W. C. Jhee, “Ensemble Size Reduction in Fraud Detection System,” *Proceedings of 2007 KMIS International Conference*, (2007), 597~602.

Sung, T. K., N. Chang, and G. Lee, “Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction,” *Journal of Management Information Systems*, Vol. 16, No.1(1999), 63~85.

Stubbs, E., *Big Data, Big Innovation*, Wiley, 2014.

Tam, K. Y., and M. Y. Kiang, “Managerial Applications of Neural Networks: The Case of Bankruptcy Predictions,” *Management Science*, Vol.38, No.1(1992), 926~947.

Wi, T. -S. and S. -K. Kwon, “Transaction Practices Reform in the Wholesale Markets for Strengthening the Competition Power,” *Korean Journal of Food Marketing Economics*, Vol.23, No.3(2006), 113~144.

Wi, T. -S. and S. -K. Kwon, “Reorganization of the Agricultural Wholesale Market,” *Korean Journal of Food Marketing Economics*, Vol.26, No.3(2009), 75~93.

Abstract

Detection of Phantom Transaction using Data Mining: The Case of Agricultural Product Wholesale Market

With the rapid evolution of technology, the size, number, and the type of databases has increased concomitantly, so data mining approaches face many challenging applications from databases. One such application is discovery of fraud patterns from agricultural product wholesale transaction instances.

The agricultural product wholesale market in Korea is huge, and vast numbers of transactions have been made every day. The demand for agricultural products continues to grow, and the use of electronic auction systems raises the efficiency of operations of wholesale market. Certainly, the number of unusual transactions is also assumed to be increased in proportion to the trading amount, where an unusual transaction is often the first sign of fraud. However, it is very difficult to identify and detect these transactions and the corresponding fraud occurred in agricultural product wholesale market because the types of fraud are more intelligent than ever before. The fraud can be detected by verifying the overall transaction records manually, but it requires significant amount of human resources, and ultimately is not a practical approach. Frauds also can be revealed by victim's report or complaint. But there are usually no victims in the agricultural product wholesale frauds because they are committed by collusion of an auction company and an intermediary wholesaler. Nevertheless, it is required to monitor transaction records continuously and to make an effort to prevent any fraud, because the fraud not only disturbs the fair trade order of the market but also reduces the credibility of the market rapidly. Applying data mining to such an environment is very useful since it can discover unknown fraud patterns or features from a large volume of transaction data properly.

The objective of this research is to empirically investigate the factors necessary to detect fraud transactions in an agricultural product wholesale market by developing a data mining based fraud detection

Seon Ah Lee* · Namsik Chang**

* Technology Division, WAVUS Co., Ltd.
** Corresponding Author: Namsik Chang
College of Business Administration, University of Seoul
163, Seoulsiripdae-ro, Dongdaemun-gu, Seoul 130-743, Korea
Tel: +82-2-6490-2228, Fax: +82-2-6490-2219, E-mail: nchang@uos.ac.kr

model. One of major frauds is the phantom transaction, which is a colluding transaction by the seller(auction company or forwarder) and buyer(intermediary wholesaler) to commit the fraud transaction. They pretend to fulfill the transaction by recording false data in the online transaction processing system without actually selling products, and the seller receives money from the buyer. This leads to the overstatement of sales performance and illegal money transfers, which reduces the credibility of market. This paper reviews the environment of wholesale market such as types of transactions, roles of participants of the market, and various types and characteristics of frauds, and introduces the whole process of developing the phantom transaction detection model. The process consists of the following 4 modules: (1) Data cleaning and standardization (2) Statistical data analysis such as distribution and correlation analysis, (3) Construction of classification model using decision-tree induction approach, (4) Verification of the model in terms of hit ratio. We collected real data from 6 associations of agricultural producers in metropolitan markets. Final model with a decision-tree induction approach revealed that monthly average trading price of item offered by forwarders is a key variable in detecting the phantom transaction. The verification procedure also confirmed the suitability of the results. However, even though the performance of the results of this research is satisfactory, sensitive issues are still remained for improving classification accuracy and conciseness of rules. One such issue is the robustness of data mining model. Data mining is very much data-oriented, so data mining models tend to be very sensitive to changes of data or situations. Thus, it is evident that this non-robustness of data mining model requires continuous remodeling as data or situation changes.

We hope that this paper suggest valuable guideline to organizations and companies that consider introducing or constructing a fraud detection model in the future.

Key Words : agricultural product wholesale market, phantom transaction, fraud detection, data mining, decision-tree induction approach

Received : November 25, 2014 Revised : January 6, 2015 Accepted : January 14, 2015
Type of Submission : Normal Track Corresponding Author : Namsik Chang

저 자 소 개

이 선 아

서울시립대학교에서 경영학 학사 및 석사 학위를 취득하였으며, 현재 주식회사 웨이버스 기술본부에 재직 중이다. 아모레퍼시픽, 롯데마트 등의 GIS 구축, 행정안전부의 차세대 전자인사관리시스템 구축 등 다수의 프로젝트에 참여하였다. 주요 관심분야는 데이터마이닝과 OLAP을 이용한 빅데이터 분석, CRM 등이다.

장 남 식

University of Missouri에서 경영학 석사, 그리고 University of Arizona에서 경영정보시스템을 전공하여 경영학 박사학위를 취득하였으며, LG-EDS시스템에서 IT 컨설턴트로 근무한 바 있다. 현재 서울시립대학교 경영대학 경영학부에 재직 중이다. 삼성, LG, 국민, 현대 카드 및 농협, 외환은행, 부산은행 등의 CRM 프로젝트를 수행하였으며, Journal of MIS, Decision Support Systems, Informs Journal on Computing, 경영정보학연구, Information Systems Review 등에 논문을 게재하였다. 주요 관심분야는 데이터웨어하우스 시스템 설계 및 구축, 데이터마이닝을 이용한 각종 산업별 빅데이터 분석 및 활용, 그리고 이들과 CRM과의 효과적인 접목 방안 연구 등이다.