Ensemble Size Reduction in Fraud Detection System
축소된 앙상블에 의한 부정행위 적발 모형

Youngmi Song, Whankyu Han and Won Chul Jhee[a]
송 영미, 한 완규, 지 원철

[a] Department of Industrial &Information Engineering, School of Information & Computer Engineering
Hong Ik University, 72-1Sangsu-dong Mapo-ku, Seoul 121-791, Korea
Tel: +82-2-320-1684, Fax: +82-2- 336-1130, E-mail:jhee@wow.hongik.ac.kr

Abstract

데이터 마이닝 분야에서 앙상블 모형의 유용성은 널리 인정되고 있다. 앙상블을 구성하는 단위모형들 사이의 다양성이 보장되는 경우, 최종 모형의 정확성 및 안정성이 향상되기 때문이다. 하지만, 얼마나 많은 단위 모형들이 어떤 방식으로 결합되어야 하는가에 대해서는 아직도 더 많은 연구가 필요하다.

본 연구에서는 신용카드 부정사용 유형 중 하나인 현금불법융통 문제에 대해 앙상블 모형의 유용성을 검증하고자 한다. 부정행위 적발 모형은 전형적인 분류 문제의 한 유형이나, 클래스간 불균형이 매우 심하다는 특징이 있다. 따라서, 현금불법융통 문제에 적합한 다양성(Diversity) 척도를 개발하여 최소한의 단위모형들로 앙상블 모형을 구성하는 방안을 제시하였다. 축소된 앙상블 모형이 많은 수의 모형을 결합한 앙상블 모형과 거의 같은 정확성 및 안정성을 보임을 국내 신용카드사의 실제 자료를 사용하여 입증하였다.

Keywords

Data Mining, Ensemble Model, Diversity Measure, Fraud Detection

1. 서론

예측모형의 일반화 능력을 제고시키기 위한 앙상블(Ensemble) 또는 위원회(Committee) 학습 방법은 지난 십여 년 동안 데이터마이닝 분야에서 많은 관심을 받아 왔다. 앙상블은 다수의 기본 모형(Base Model)들을 결합하여 예측의 정확성 및 안정성을 높이는 것으로, 첫째 앙상블을 구성하는 기본 모형들을 어떻게 얼마나 생성할 것인가 하는 문제와 둘째 생성된 기본 모형들의 예측치들을 어떻게 결합하여 예측성능을 개선할 것인가의 두 문제로 나누어 진다.

앙상블에 대한 과거 연구들은 앙상블을 구성하는 기본 모형들이 다양하면서 최소한의 정확성을 가질 경우에 효과적이었음을 보였으며, 기본 모형들의 다양성을 확보하기 위한 여러 시도가 있었다. 하지만 기본 모형들 사이의 다양성과 정확성은 서로 상충관계에 놓일 경우가 많으므로 앙상블 구성할 때 양자에 대한 적절한 고려가 필요하다.

다양성을 고려한 앙상블에 관한 연구의 대부분은 UCI 레포지터리의 벤치마킹 자료를 이용한 것이다. 물론, 연구의 객관성을 확보하기 위하여 벤치마킹 자료를 사용하는 것은 바람직하지만, 기존 연구에 사용된 데이터의 규모가 매우 작다는 사실을 고려할 때, 해당 연구 결과가 현실세계에서도 그대로 적용될 수 있다는 보장은 없다. 특히, 대용량 데이터로부터 유용한 지식을 발견한다는 데이터 마이닝 관점에서 볼 때 현실의 대용량 데이터를 이용한 연구가 아직도 필요하다.

최근 신용카드사에서 신용카드 부정행위 적발시스템(Fraud Detection System, FDS)은 필수적인 시스템으로 인식되고 있으며, FDS 에서 부정행위 적발 모형(Fraud Detection Model, FDM)은 핵심적 구성요소로서 앙상블을 이용한 모델링 기법들이 시도되고 있다. 신용카드를 사용한 부정행위의 유형은 분실• 도난, 위변조, 명의도용, 정보유출, 현금불법유통, 매출표 유통, 위장 가맹점 등으로 분류할 수 있으며, FDM 은 분실• 도난과 위변조를 중심으로 모델링이 이루어져 왔다. 신용카드의 FDM 도 전형적인 분류 문제의 하나이지만 분류 대상이 되는 클래스들간의 불균형이 매우 심하다는 특징이 있다.

본 연구에서는 신용카드 부정행위 중 현금불법유통(속칭 카드깡) 적발 모형을 개발하는 과정에서 1)신용카드사의 실제 상황을 최대한 반영한 자료를 이용하고, 2) 앙상블 기법을 사용함에 있어 다양성 척도를 고려하여 앙상블에 포함되는 기본 모형들의 수를 최소화할 수 있는가를 검증하고자 한다. 현금불법융통이란 신용카드 가맹점이 실제 물품의 판매나 용역의 제공 없이 신용카드에 의한 거래를 가장하여 매출을 발생시키고, 허위매출금을 카드회사에 청구하는 방법으로 현금을 융통하는 행위로, 여신전문금융업법 제 70 조에 의거하여 3 년 이하의 징역 또는 2 천만 원 이하의 벌금을, 이용고객은

금융질서 문란 자로 7 년간 금융거래상의 제한을 받게 된다. 불법현금유통은 대부업자들이 재무적 위험을 카드사에게 전가시키는 행위로 이용 고객들이 악성연체 상태에 빠질 확률이 매우 높다는 점에서 카드사 입장에서는 반드시 방지하여야 할 부정행위이다.

본 논문의 구성은 첫째, 신용카드 FDM 에 적절한 정확성과 다양성 척도에 대해 논의한다. 둘째, 국내 신용카드사로부터 입수한 자료로부터 FDM 개발을 위한 학습자료의 구성에 대해 설명한다. 셋째, 로짓 회귀모형, 의사결정수 및 신경망을 이용한 앙상블 구성에 대해 실험을 설명한 후, 마지막으로 다양성 척도를 사용함으로써 상대적으로 작은 수의 기본 모형들로 구성한 앙상블- 즉 축소된 앙상블 모형의 성능에 대해 논의한다.

2. 앙상블과 다양성

앙상블을 구성함에 있어 기본 모형의 다양성을 확보하는 방법은 크게 세가지로 요약할 수 있다. 첫째는 기본 모형에 사용되는 입력변수를 모형 별로 다르게 구성하는 것이고, 둘째는 기본 모형의 학습 파라메터들을 다양하게 하는 것이다. 예를 들어, 신경망의 경우 기본 모형들의 신경망 구조나 초기 가중치를 다르게 할 수 있고, 의사결정수의 경우 분리(Splitting) 기준을 다르게 할 수 있다. 세 번째는 기본 모형들을 서로 다른 학습집합에 대해 학습시키는 것이다.

Boot-strap Sample 을 사용하는 Bagging 이나 기 학습된 모형의 성능을 감안하여 적응적으로 학습자료를 구성하는 Boosting 은 학습 자료의 다양성으로 앙상블을 구성하는 대표적인 방법들이다. 하지만, 전체 학습 자료의 양이 충분하다면 기본 모형 별로 완전히 다른 학습 자료를 구성하여도 된다. 또, 앙상블을 구성하는 기본 모형들에 모두 동일한 학습 알고리즘을 사용할 수 도 있지만, 서로 다른 학습 알고리즘을 사용하여 다양성을 확보하는 이질적(heterogeneous) 앙상블을 구성할 수 도 있다.

앙상블을 구성하는 기본 모형들이 학습되면 기본 모형들의 출력값들을 결합하는 방법으로는 Majority Voting 과 단순 평균 등이 많이 사용되었으며, 최근에는 기본 모형들의 출력값으로 상위 예측 모형을 구성하는 메타모형에 대한 연구가 있다. 기본 모형의 출력값들을 결합할 때, 학습된 모든 기본 모형들을 사용하여 왔으나, 최근 다양성 척도를 고려하여 앙상블의 성능향상에 도움이 되는 기본 모형들만 포함시키는 연구들이 있었다.

본 연구에서는 신용카드 불법현금융통 적발을 위한 FDM 을 개발함에 있어 앙상블의 성능향상에 도움이 되는 기본 모형들만 선별하여 앙상블을 구성하고자 하며, 선별기준으로는 정확성과 다양성을 동시에 고려하고자 한다. 따라서, 최종 앙상블에 포함될 기본 모형의 선별 기준인 공헌도를 다음과 같이 사용하였다.

공헌도 = α X 정확성 + (1-α)X 다양성

이러한 척도는 Optiz[]가 처음 사용하였으며, 앙상블을 구성할 때 공헌도가 큰 순서대로 기본 모형의 선택하게 된다. α 는 정확성과 다양성의 반영비율로서 0 과 1 사이의 값을 갖는다.

공헌도를 구성하는 정확성 척도로서는 Weighted Efficiency(WE)를 사용하였다. WE 를 <표 1>과 같은 오분류표를 이용하여 정의하면 다음과 같다.

<표 1> FDM의 오분류표

		예측		合
		정상(0)	사고(1)	
실제	정상(0)	*N00*	*N01*	*N0*
	사고(1)	*N10*	*N11*	*N1*

WE = (OC)*(FC)*(FPC)

where OC = (N00+N11)/(N0+N!)

FC = N11/N1

FPC = N11/(N01+N11)

신용카드 현금불법융통 적발과 같은 FDM 문제의 성능 측정은 일반적인 분류 문제의 모형 성능 측정과 다른 특성을 가진다. 정상 건의 수(N0)가 사고건에 비해 압도적으로 많기 때문에 특이도(Specificity)가 대부분의 모델에서 매우 높은 수치를 기록해 전체 정분류율을 왜곡시키고, 오분류비 문제 때문에 민감도(Sensitivity) 역시 모형 성능 척도로 활용하기 어렵다. 목표변수의 Class 별(정상, 사고) 빈도수 차이가 심각하게 큰 경우에는 N11 을 최대화하면서, N01 을 최소화하는 모형이 가장 좋은 모형이다. 따라서, 제 2 종 오류(사고 건을 정상 건이라고 잘못 예측)와 관련된 FC(fraud Classification)와 제 1 종 오류(정상 건을 사고 건이라고 잘못 예측)와 관련된 FPC(Fraud Prediction Classification)를 동시에 고려할 수 있는 WE 는 FDM 에 적절한 척도이다.

Kuncheva et al 은 10 개의 다양성 척도에 대해 서로의 관계 및 앙상블의 정확도에 미치는 영향을 분석하였지만, 다양성의 고려가 정확도 향상에 기여하는 것은 사실이지만 어떤 다양성 척도가 가장 바람직한가에 대한 결론은 내리지 못했다. 따라서 본 연구에서는 가장 일반적으로 사용된 다양성 척도인 분산(variance) 개념을 사용하기로 한다. 기본 모형의 다양성이란 같은 입력 값이 들어갔을 때 서로 다른 출력 값이 나올수록 높아지는 것으로, 모형의 다양성을 측정하기 위하여 Brodley 가

1996 년에 처음 사용한 Ambiguity Measure 를 사용하였다. Ambiguity Measure 는 같은 입력 값이 들어갔을 때, 전체 앙상블 모형과, 다양성을 측정하고자 하는 i 번째 모형과의 출력 값 차이를 측정하여 구한다. 따라서, 다양성 척도 (Diversity, Div)는 다음과 같이 정리할 수 있다.

$$\mathrm{Div}_i^{Full} = \sum_{k=1}^{n} \{ f_i(x_k) - \overline{f(x_k)} \}^2$$

where $i = 1 \cdots n$ (n : 기본모형의 수)

$k = 1 \cdots m$ (m : 입력자료의 수)

$f_i(x_k)$: i 번째 모형의 k번째 입력의 출력값

$\overline{f(x_k)}$: 앙상블모형의 k번째 입력의 출력 값

FDM 문제에서는 목표변수의 class 별 빈도수가 절대적으로 차이가 나는 점을 고려하여 다양성 척도를 세분하였다. 즉, Ambiguity Measure 를 목표변수 class 별로 측정하면 입력자료 전체의 Div 와 정상건의 Div, 사고건의 Div 의 3 가지로 분류할 수 있다. 하지만, 전체 학습자료 중 정상건의 비율이 사고 건에 비해 절대적으로 많으므로 정상건의 Div 값은 전체 자료의 Div 값과 거의 비슷하였기 때문에 제외하였다. 또, 두 가지 다양성 척도와 정확성 척도인 WE 값 사이의 상대 비교가 불가능한 문제를 해결하기 위하여 각 기본 모형 별 측정값을 전체 기본 모형의 측정값의 합계로 나누어 표준화하였다.

따라서, 공헌도를 측정하기 위한 정확성 척도(WE)와 다양성 척도(Div)는 다음과 같이 정의하였다.

$$\mathrm{Std\_WE}_i = \frac{WE_i}{\sum_{j=1}^{n} WE} \quad \cdots\cdots (1)$$

$$\mathrm{Std\_Div\_all}_i = \frac{Diversity_i^{full}}{\sum_{i=1}^{n} Diversity_i^{full}} \quad \cdots\cdots (2)$$

$$Std\_Div\_1_i : \frac{Diversity_i^{Fraud}}{\sum_{j=1}^{n} Diversity_j^{Fraud}} \quad \cdots\cdots (3)$$

$$\mathrm{Std\_Div\_all2}_i : \frac{Std\_Div\_1_i + Std\_Div\_0_i}{2} \quad \cdots\cdots (4)$$

where. $j = 1 \cdots n$ (n : 단일 모형의 수)

3. 실험 설계

3.1. 실험 자료

본 연구에 사용된 자료는 국내 카드사의 거래 자료인 승인자료들이다. 신용카드 현금불법융통 거래의 특징은 두 가지로 정리할 수 있다. 첫째, 높은 금액대, 높은 할부 개월 수에서 발생하는 경향이 있다. 고액의 승인을 발생시킨 뒤, 할부로 천천히 청구되도록 하기 때문이다. 둘째, 월 평균 사고율이 매우 낮다. 대용량의 신용카드 승인 건에 비해 사고 건은 적기 때문에 목표 변수의 Class 별 빈도 수 차이가 심하다.

따라서, 2005 년 10 월부터 12 월까지 3 달 동안의 승인자료 중 20 만원 이상 할부 3 개월 이상인 170 만 건으로부터 학습자료를 구성하였는데, 정상대 사고 비율은 950:1 이었다. 학습된 모형에 대한 테스트 자료로는 2006 년 1 월 의 48 만 건을 1 차로 사용하였으며, 모형의 안정성을 확인하기 위한 2 차 테스트 자료로 2006 년 08 월 데이터 37 만 건을 사용하였다. 1 차 테스트 자료의 정상 대 사고 비율은 1 차는 1245:1, 2 차는 676:1 이었다.

학습자료는 현금불법융통 적발 사고 자료와 승인 데이터, 회원/카드/가맹점정보 등을 이용한 입력 변수로 구성되었다. 회원 승인 데이터와 회원속성/카드속성/가맹점속성 데이터 및 Profile 집계를 이용하여 기본 변수를 생성하였고, 각 변수 별 데이터 분석을 실시하여 범주화/변수변환/그룹화/interaction 등의 변수를 추가 가공하여 사용하였다.

3.2. 기본 모형의 학습

학습된 기본 모형의 다양성을 확보하기 위하여 다음과 같이 세가지 방법을 사용하였다.

첫째, 학습 대상 샘플을 다양화하였다. 목표변수의 class 별 빈도 차이가 현격할 경우 학습 데이터의 class 별 빈도 차이를 조절해주지 않으면 모형의 학습이 불가능하기 때문에 샘플의 크기와 목표변수의 class 별 비율을 결정하기 위해 실험을 실시하였다. 다양한 크기와 다양한 비율의 샘플을 여러 개 추출하여 Decision Tree 와 Logit Regression 을 이용한 파일루트 테스트를 실시하였다. 각 샘플 내 정상건의 수를 25 만에서 55 만까지 변화를 시켰을 경우 학습된 모형의 성능에 뚜렷한 차이는 보이지 않았으나, 전체 정상건의 패턴을 잘 반영하며 다른 샘플 크기보다 성능이 좋았던 40 만 건을 최종 선택하였다. 정상 대 사고 비를 5:1 에서 20:1 까지 실험해 보았을 때, 8:1 ~ 10:1 이 되도록 조정을 해 주는 것이 가장 모형의 성능이 좋았다. 따라서, 정상건의 경우 자료가 충분하므로 임의추출에 의해 사고건은 boot-strapping 에 의해 정상 대 사고 비율을 맞추었으면 총 80 개의 학습용 샘플을 생성하였다.

둘째, 기본 모형에 사용된 학습 알고리즘을 다양화하기 위하여 로짓분석, 의사결정수, 신경망의 세가지를 사용하였다. 생성된 샘플 80 개는 두 개씩

짝을 지어 하나는 학습용(Train)으로 다른 하나는 검증용 (Validation)으로 사용하였다. 따라서, 로짓분석과 의사결정수의 경우는 40 개씩의 기본 모형이 학습되었다. 하지만 신경망의 경우 학습시간을 고려하여 10 개만 학습하였다.
셋째, 각 모형 별 입력 변수 조합을 다르게 주었다. 입력 변수 조합을 결정할 때는 각 학습 알고리즘 별로 다른 기준을 적용하였다. 의사결정수는 자체적으로 변수선택 기능이 있으므로 모든 변수를 입력 변수로 선정하여 학습하였고, 로짓분석의 경우 Stepwise 절차(유의수준 5%)에 의해 학습하였다. 신경망의 경우 변수선택을 위한 절차가 없으므로 의사결정수와 로짓분석에서 선택된 변수들을 정리한 후, 인위적으로 조정하였다.

3.3. 축소된 앙상블

본 연구의 목적은 실험을 통해 다음의 세가지를 확인하는 것이다. 첫째, 신용카드 현금불법융통 적발 모형에 앙상블 사이즈 축소 방안을 적용한 경우, 효과가 있는지를 검증해본다. 다량의 단일 모형을 학습한 뒤, 모든 단일 모형을 결합한 전체 앙상블 모형(Full Ensemble Model)과 기본 모형을 선택하여 결합한 축소된 앙상블 모형(Pruned Ensemble Model)의 성능을 비교하는 것이다. 둘째, 목표변수의 Class 별 빈도 차이가 현격한 경우, 사고 건을 다양하게 예측하는 단일 모형들로 앙상블을 구성하는 것이 좋은지 정상 건을 다양하게 예측하는 단일 모형들로 앙상블을 구성하는 것이 좋은지를 다양성 척도를 다르게 함으로써 실험하였다. 셋째, 동질적 학습(Homogeneous Learning)과 이질적 학습(Heterogeneous Learning) 각각에 앙상블 사이즈 축소를 실험하여 효과를 검증하였다.
2 장에서 설명한 공헌도에 따라 앙상블에 포함될 기본 모형들을 선정하기 위하여 정확성(WE) 및 다양성(Div) 척도와 가중치 α 을 이용하여 <표 2>와 같이 13 가지 선택 기준을 결정하였다. 실험은 13 가지 선택 기준에 따라 공헌도가 높은 순서대로 상위 N 개의 기본 모형으로 축소된 앙상블을 구성하는 방식으로 진행하였다. N= 5, 10, 15, 20 의 네 가지 조합 군을 실험에 사용함으로써 총 52 가지의 축소된 앙상블 후보 조합을 생성하였다. N=5, 10, 15, 20 에 대해 Top5, Top10, Top15, Top20 으로 각각 표시하기로 한다.
앙상블을 구성하기 위해 기본 모형들의 출력값을 결합하는 방법으로는 단순평균(Simple Average)을 사용하였다. 앙상블을 이용한 분류 문제에 있어 많이 사용되는 Majority Voting 을 사용하지 않은 이유는 파일루트 테스트 결과 단순평균이 FDM 에서는 더 좋은 결과를 보였기 때문이다.

<표 2> 앙상블 구성원 선택 기준

기준	Accuracy Measure	Diversity Measure	α
1	(1) Std_WE	-	1
2	-	(2) Std_Div_all	0
3	-	(3) Std_Div_1	0
4	-	(4) Std_Div_all2	0
5	(1) Std_WE	(2) Std_Div_all	0.25
6	(1) Std_WE	(2) Std_Div_all	0.5
7	(1) Std_WE	(2) Std_Div_all	0.75
8	(1) Std_WE	(3) Std_Div_1	0.25
9	(1) Std_WE	(3) Std_Div_1	0.5
10	(1) Std_WE	(3) Std_Div_1	0.75
11	(1) Std_WE	(4) Std_Div_all2	0.25
12	(1) Std_WE	(4) Std_Div_all2	0.5
13	(1) Std_WE	(4) Std_Div_all2	0.75

4. 실험 결과

각 기본 모형 및 앙상블 모형의 성능을 비교하는 척도로서 LIFT 값을 사용하였다. 테스트 데이터에 스코어링을 실시하여 스코어가 높은 순으로 데이터를 정렬하고 20 등분을 한 뒤 1 등급에서의 LIFT 값을 비교하였다. 1 등급에서의 LIFT 값을 사용한 이유는, 정상 건의 수가 매우 많고 사고 건의 수가 매우 적어 2 등급 이하의 등급에서의 성능 평가는 큰 의미가 없기 때문이다. (1 차 테스트에서의 1 등급 LIFT 값), (2 차 테스트에서의 1 등급 LIFT 값)처럼 쉼표를 사이에 두고 구분하여, 연속적으로 성능을 서술하였다.
기본 모형 성능은 <표 3>에 요약하였다. 90 개 전체 단일모형 성능의 평균은 14.0, 12.9 이고, 분산은 1.5, 1.6 이다. 90 개 기본모형을 모두 선택하여 만든 전체 앙상블 모형의 성능은 16.4, 15.3 으로, 앙상블을 구성할 경우 성능이 향상됨을 뚜렷하게 알 수 있었다.

<표 3> 기본 모형 성능 요약

개요		단일모형 평균		단일모형 분산		앙상블 성능	
종류	개수	1 차 TEST	2 차 TEST	1 차 TEST	2 차 TEST	1 차 TEST	2 차 TEST
NN	10	14.7	13.4	1.3	1.0	16.9	15.1
TR	40	14.0	12.8	0.8	1.5	16.4	14.9
REG	40	13.8	12.8	2.1	1.8	15.2	14.1
전체	90	***14.0***	***12.9***	1.5	1.6	***16.4***	***15.3***

<표 4>는 의사결정수(TR)만으로 동질적 앙상블을 구성한 실험 결과를 요약한 것이다. TR 40 개 모형으로 구성한 TR 전체 앙상블 모형보다 성능이 좋은 축소된 앙상블 모형이 존재한다. <표 4>에서 굵은 글씨체로 표시한 조합이다. Best Case 는 기준 6 의 Top10, Top15, Top20 인데, 다양성

척도로 (2)Std_Div_all를 사용하고 α =0.5를 적용한 것이다. 이는 정상 건을 다양하게 예측하는 단일 모형들로 앙상블을 구성하는 것이 사고 건을 다양하게 예측하는 단일 모형들로 앙상블을 구성하는 것보다 효과적이며, 정확성과 다양성을 함께 반영하여 기본 모형들을 선택하는 것이 효과적임을 보여준다. TR 단일 모형보다 성능이 좋지 않은 축소된 앙상블 모형이 존재하므로 앙상블 구성원과 크기를 잘못 선택할 경우, 앙상블 모형을 구성하지 않는 것보다 못하다는 것을 알 수 있었다. 또한 구성원이 20 개 이상인 모든 앙상블 모형은 단일 모형의 최고 성능보다 뛰어난 모형을 보였는데 이는 앙상블 크기가 일정 개수 이상이 될 때에는 구성원을 어떻게 선택하더라도 단일 모형보다 성능이 좋아진다는 것을 나타낸다.

<표 4> 동질적 앙상블(TR40)의 성능

	1 차 TEST				2 차 TEST			
	top5	top10	top15	top20	top5	top10	top15	top20
기준 1	15.3	15.5	15.4	15.4	14.4	14.6	14.7	14.8
기준 2	15.0	16.3	***16.4***	***16.4***	12.3	13.8	14.2	14.5
기준 3	15.0	16.3	***16.4***	***16.4***	12.3	13.8	14.2	14.5
기준 4	15.0	15.8	16.1	16.2	12.3	13.6	13.7	14.0
기준 5	15.0	16.3	***16.5***	***16.4***	12.3	14.1	14.5	14.5
기준 6	15.0	***16.7***	***16.4***	***16.4***	12.3	***14.9***	***14.9***	***15.0***
기준 7	15.5	16.1	16.1	***16.4***	14.5	***15.0***	***15.0***	***15.0***
기준 8	15.5	15.9	16.2	16.3	13.8	13.6	13.8	14.3
기준 9	15.5	15.9	16.2	16.3	14.1	13.6	13.8	14.3
기준 10	15.5	15.5	15.6	15.6	14.7	14.7	14.8	***15.0***
기준 11	15.0	15.5	16.2	16.3	12.3	13.6	14.1	14.2
기준 12	15.0	16.2	16.3	16.2	12.3	14.3	14.8	***15.0***
기준 13	15.5	15.9	15.8	15.8	14.5	14.8	***15.0***	***15.1***
TR 앙상블(40)	***16.4***				***14.9***			
단일모형 최고성능(TR31)	15.4				14.1			

<표 5>는 로짓분석(REG)만으로 동질적 앙상블을 구성한 실험결과이다. REG 40 개 모형으로 구성한 REG 전체 앙상블 모형보다 성능이 좋은 축소된 앙상블 모형이 존재한다. <표 5>에서 굵은 글씨체로 표시한 조합이다. Best Case 는 기준 7, 기준 10, 기준 13 의 Top10 인데, α =0.75 를 적용했다는 공통점이 있다. 이는 정확성과 다양성 척도의 반영 비율을 3:1 정도로 잡는 것이 가장 효율적이라는 의미이다. 한편 다양성 에 중점을 둔, 즉 α <=0.5 인 모든 축소된 앙상블 모형의 성능은 BEST REG 단일 모형보다 좋지 않았다. REG 의 경우 TR 나 NN 보다 모형들의 평균 성능이 낮고 분산이 크기 때문에, 다양성보다는 정확도에 더욱 비중을 두어 앙상블을 구성하는 것이 효율적이라는 것을 추측할 수 있다. <표 5>에서 특이한 사실을 발견할 수 있는데, Top15의 성능이 Top5, Top10, Top20 보다 낮다는 사실이다. 앙상블 구성원의 개별 성능이 낮고, 분산도 높은 상태에서 축소된 앙상블을 구성하는 경우에는 특히 주의를 기울여야 한다는 사실을 보여준다.

<표 5> 동질적 앙상블(REG40)의 성능

	1 차 TEST				2 차 TEST			
	top5	top10	top15	top20	top5	top10	top15	top20
기준 1	***15.3***	***15.4***	15.1	15.1	14.1	***14.4***	14.0	14.1
기준 2	12.4	13.0	14.5	***15.6***	11.7	12.0	13.4	14.0
기준 3	12.3	13.0	14.5	***15.2***	11.9	12.0	13.5	13.9
기준 4	12.3	13.0	14.5	***15.5***	11.9	12.0	13.3	13.9
기준 5	12.7	13.0	14.7	***15.6***	11.8	12.0	13.4	14.0
기준 6	12.8	13.0	14.9	***15.5***	12.0	12.0	13.7	14.0
기준 7	15.0	***15.4***	15.1	***15.3***	14.2	14.4	14.0	14.1
기준 8	12.3	13.0	14.5	***15.5***	11.9	12.0	13.5	13.9
기준 9	12.5	13.0	14.8	***15.4***	11.8	12.0	13.5	14.1
기준 10	15.1	***15.5***	15.1	15.0	***14.4***	***14.4***	14.0	13.9
기준 11	12.3	13.0	14.6	***15.6***	11.9	12.0	13.4	13.9
기준 12	12.4	13.0	14.9	***15.4***	11.7	12.0	13.7	14.0
기준 13	***15.3***	***15.4***	15.1	***15.2***	14.1	***14.4***	14.0	14.1
REG 앙상블(40)	***15.2***				***14.1***			
단일모형 최고성능(REG104)	14.8				14.1			

<표 6>은 이질적 앙상블 구성을 실험한 결과이다. REG 40 개/TR 40 개/NN 10 개 총 90 개 기본 모형으로 구성한 전체 앙상블 모형보다 성능이 좋은 축소된 앙상블 모형이 존재한다. Best Case 는 다양성 척도로 (2)Std_Div_all 을 적용한 기준 7 과 (4)Std_Div_all2 를 적용한 기준 13 인데, α =0.75 이라는 공통점이 있다. 이는 정확성과 다양성을 반영 비율을 3:1 정도로 잡는 것이 가장 효율적이라는 의미이다. α =0 이나 α =1 인 경우의 성능은 매우 낮은데, 특히 다양성 척도 (4)Std_Div_1 를 100% 반영하여 앙상블 구성원을 선택한 기준 3 의 Top5 에서 Top20 까지의 네 가지 조합의 축소된 앙상블 모형은 모두 전체 앙상블 모형보다 낮은 성능을 보였다. 이는 FDM 문제에서 앙상블을 구성할 경우, 사고 건을 다양하게 예측하는 구성원을 선택하는 것보다, 정상 건을 다양하게 예측하는 구성원들로 구성하는 것이 좋다는 것을 의미한다.

<표 4, 5, 6>을 통하여 의사결정수와 로짓분석을 이용한 동질적 앙상블 구성에서 축소된 앙상블 효과를 확인할 수 있었으며, 신경망을 포함한 이질적 앙상블 구성에서도 축소된 앙상블은 효과가 있었다.

5. 결론

본 연구에서는 신용카드 현금불법융통 적발 모형의 개발에 있어 앙상블 사이즈 축소가 유효한가를 실험하기 위하여 국내 카드사로부터 수집한 대용량

자료를 사용하였다. 앙상블을 구성하는 기본 모형들의 다양성을 확보하기 위하여 80 개의 학습자료의 구성, 로짓분석/의사결정수/신경망의 세 가지 학습 알고리즘의 사용 및 학습시 입력변수의 다양성이 보장되도록 하였다. 또, 축소된 앙상블을 구성하는 기본 모형의 선택기준으로 Weighted Efficiency 에 의한 정확성 척도와 Ambiguity 에 의한 다양성 척도를 동시에 고려하였다. 본 연구의 실험 결과를 요약하면 다음과 같다.

첫째, 신용카드 불법현금융통(카드깡) 적발 모형 개발에 앙상블 사이즈 축소를 적용할 경우, 성능을 유지 혹은 개선하면서 모형의 복잡도를 낮출 수 있다. 둘째, 기본 모형의 생성 시 Homogeneous Learning 보다는 Heterogeneous Learning 이 더욱 효과적이다. 셋째, 정확성과 다양성을 함께 고려하여 축소된 앙상블 모형의 구성원을 선택하는 것이 효과적이다. 넷째, 목표변수의 Class 별 빈도 차이가 큰 부정행위 적발모형의 경우, 사고건을 다양하게 예측하는 구성원들 보다는 정상건을 다양하게 예측하는 구성원 들로 축소된 앙상블을 구성하는 것이 효과적이다. 마지막으로 본 논문에서는 반년이 경과한 뒤의 모형 성능 변화 추이를 살펴보았다. 축소된 앙상블이 모든 기본 모형을 사용한 앙상블과 비교하여 성은에 큰 차이를 보이지 않았다.

<표 6> 이질적 앙상블(90) 성능

	1 차 TEST				2 차 TEST			
	top5	top10	top15	top20	top5	top10	top15	top20
기준 1	16.0	16.2	16.3	16.2	***15.7***	***15.6***	***15.5***	15.3
기준 2	14.8	16.2	***16.8***	***16.6***	13.2	14.9	14.8	14.8
기준 3	14.8	15.4	16.3	16.3	13.8	14.1	14.6	14.9
기준 4	14.8	16.3	***16.5***	16.3	14.1	14.6	14.7	14.5
기준 5	15.5	16.2	***16.7***	***17.1***	14.5	14.9	***15.3***	***15.5***
기준 6	15.6	***16.5***	***17.0***	***17.3***	14.5	***15.3***	***15.6***	***15.7***
기준 7	16.3	***16.8***	***16.9***	***16.7***	***15.5***	***15.7***	***15.8***	***15.5***
기준 8	14.8	16.0	16.2	***16.5***	13.8	14.9	14.7	14.9
기준 9	15.1	***16.5***	***17.0***	***17.1***	14.8	***15.6***	***15.6***	***15.5***
기준 10	16.3	***16.6***	16.4	***16.5***	***15.5***	***15.8***	***15.4***	***15.5***
기준 11	14.6	16.2	***16.6***	***17.0***	14.6	14.9	15.0	***15.3***
기준 12	15.5	***16.6***	***16.9***	***17.1***	14.6	***15.5***	***15.8***	***15.7***
기준 13	16.3	***16.6***	***16.9***	***16.6***	***15.5***	***15.8***	***15.5***	***15.5***
Full 앙상블(90)	***16.4***					***15.3***		
단일모형 최고성능(NN84)	15.1					15.0		

6. 참 고 문 헌

[1] 조성목. (2004) "신용카드불법거래 유형 및 대응방안," 신용카드 제 30호.

[2] Zhao, Y., Gao, J., and Yang, X. (2005) "A survey of neural network ensembles," International Conference on Neural Networks and Brain.

[3] Hernandez, C., Fernandez, M., and Oritiz, M. (2003) "A comparison of ensemble methods for multilayer Feedforward networks," International Joint Conference on Neural Networks.

[4] Hernandez, C., Fernandez, M., and Oritiz, M. (2005) "New experimental ensembles of multilayer Feedforward for classification Problem," International Joint Conference on Neural Networks, 2005.

[5] Rooney, N., Patterson, D., and Nugent, C., (2004) "Reduced Ensemble Size Stacking," 16th IEEE International Conference on Tools with Artificial Intelligence.

[6] Mena, J. (2003). *Investigative Data Mining for Security and Criminal Detection.* BH, Elsevier Science.

[7] NEVES, J.C., and VIEIRA, A. (2006) "Improving Bankruptcy Prediction with Hidden Layer Learning Vector Quantization," European Accounting Association.

[8] Rooney, N., Patterson, D., and Nugent, C. (2006) "Pruning extensions to stacking," Intelligent Data Analysis, Vol. 10, pp47–66.

[9] Zenko, B., Todorovski, L., and Dzeroski, S., (2001) "A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods," Proceedings of the 2001 IEEE International Conference on Data Mining, pp 669-670.

[10] Zhou, Z, Wu, J., and Tang W., (2002) "Ensembling Neural Networks: Many Could Be Better than All," Artificial Intelligence, Vol. 137, pp.239-263.

[11] Zhou, Z.H. and Tang, W. (2003) "Selective Ensemble of Decision Trees," RSFDGrC 2003, pp476-483

[12] Tsymbal, A., Puuronen, S., and Patterson, D. (2003) "Ensemble feature selection with the simple Bayesian classification," Information Fusion Vol. 4, pp87-100.

[13] Opitz, D. "Feature Selection for Ensembles," (1999) Proc. 16th National Conf. on Artificial Intelligence, pp. 379-384.