韓國IT서비스學會誌
第11卷 第4號
2012年 12月, pp.233-250
http://dx.doi.org/10.9716/KITS.2012.11.4.233

신용카드 대손회원 예측을 위한 SVM 모형*

김진우** · 지원철***

Credit Card Bad Debt Prediction Model based on Support Vector Machine*

Jin Woo Kim** · Won Chul Jhee***

■ Abstract ■

In this paper, credit card delinquency means the possibility of occurring bad debt within the certain near future from the normal accounts that have no debt and the problem is to predict, on the monthly basis, the occurrence of delinquency 3 months in advance. This prediction is typical binary classification problem but suffers from the issue of data imbalance that means the instances of target class is very few.

For the effective prediction of bad debt occurrence, Support Vector Machine (SVM) with kernel trick is adopted using credit card usage and payment patterns as its inputs. SVM is widely accepted in the data mining society because of its prediction accuracy and no fear of overfitting. However, it is known that SVM has the limitation in its ability to processing the large-scale data. To resolve the difficulties in applying SVM to bad debt occurrence prediction, two stage clustering is suggested as an effective data reduction method and ensembles of SVM models are also adopted to mitigate the difficulty due to data imbalance intrinsic to the target problem of this paper.

In the experiments with the real world data from one of the major domestic credit card companies, the suggested approach reveals the superior prediction accuracy to the traditional data mining approaches that use neural networks, decision trees or logistics regressions. SVM ensemble model learned from T2 training set shows the best prediction results among the alternatives considered and it is noteworthy that the performance of neural networks with T2 is better than that of SVM with T1. These results prove that the suggested approach is very effective for both SVM training and the classification problem of data imbalance.

Keyword : Credit Card Delinquency, Bad Debt, SVM(Support Vector Machine), Data Imbalance, Ensemble Model, Data Mining

논문투고일 : 2012년 10월 22일 논문수정완료일 : 2012년 12월 03일 논문게재확정일 : 2012년 12월 06일
* 이 논문은 홍익대학교 교내연구비에 의하여 지원되었음.
** 홍익대학교 산업공학과 박사과정
*** 홍익대학교 산업공학과 교수

1. 서 론

국내 신용카드 시장규모는 IMF 사태 이후 급격히 팽창하였으며, 2003년 전후의 카드대란 시 대규모 연체채권의 발생으로 침체기를 경험하였으나, 이 후 견실한 성장을 지속해왔다. 카드 대란 이후 국내 신용카드사들은 위험 관리의 중요성을 인식하고 선진 경영기법의 도입 노력을 해왔으며, 채권관리 분야에서도 스코어링 모형을 사용한 체계적인 채권회수시스템을 구축 사용하고 있다.

<표 1>에서 볼 수 있듯이 카드 대란 이후 국내 신용카드 이용 실적은 현금대출이 감소하고 신용판매 비중이 꾸준히 높아지는 건실한 매출 구조를 보여주고 있으나, 글로벌 금융위기를 전후로 카드론 이용실적이 급증하고 있어 채권관리에 안심할 수 있는 상황은 아니다[1].

〈표 1〉 국내 신용카드 이용실적

(단위 : 조 원)

		연도					
		2004	2005	2006	2007	2008	2009
신용판매	일시불	188.0	213.0	229.9	254.5	287.2	300.9
	할부	41.9	45.2	49.0	57.6	69.0	71.7
		229.9	258.2	278.9	312.1	356.2	372.6
현금대출	현금서비스	127.6	105.2	91.6	85.8	88.8	81.5
	카드론	10.5	8.0	11.8	16.0	19.2	18.0
		138.1	113.2	103.4	101.8	108.0	99.5
합계		368.0	371.4	382.3	413.9	464.2	472.1

주) 국내 회원의 국내/해외 이용실적 합계 기준(직불카드 제외).

대부분의 국내 신용카드사들이 도입한 채권 스코어링 시스템은 채권회수 모형(Collection Scoring Model)을 이용하여 이미 연체된 채권의 효율적 회수 방안을 구현하는데 이용되고 있다. 하지만, 기 발생된 연체채권 회수 위주의 채권 관리 방식은 경기침체기에 급증하는 연체채권에 대한 대비책이 될 수 없으며, 신용카드사의 주요 경영지표인 정상입금율을 개선할 수 있는 수단을 제공하지 못한다. 따라서 선제적 채권관리가 되기 위해서는 사전에 신용카드 회원의 연체가능성을 예측하여 연체가 발생하기 전에 필요한 조치를 취해야 한다. 회원의 연체가능성 예측은 연체 채권의 회수가능성 예측 보다 정확도가 떨어질 수 있다는 어려움은 있지만, 신용카드사의 경영상태가 금융산업의 선행 지표가 되고 있는 점을 감안하면 신용카드사의 대량 연체회원 발생은 한 국가의 금융경색으로 이어질 가능성이 높으므로 보다 적극적인 채권관리 수단으로써 연체가능성 예측모형의 개발 및 운영 노하우를 축적할 필요가 있다. 특히, 장기 악성 화된 고액의 채권들은 결국 대손 처리되어 카드사의 수익성을 악화시키므로 신용카드사들의 특별한 관심이 필요하다.

본 논문에서는 신용카드 회원의 연체가능성을 사전 예측하는 모형의 하나로 정상회원의 대손처리 가능성을 카드 사용행태 및 입금 실적 등을 사용하여 개발하였다. 본 논문의 연체예측 모형은 특정 월에 연체를 시작하여 향후 대손 처리될 회원들을 예측한다는 점에서 단순히 향후 부실화에 의한 연체가능성이 있는 회원을 예측하는 기존의 신용평가 모형들과는 차이가 있다. 따라서 본 논문의 모형 개발에는 예측의 정확도가 매우 높은 데이터마이닝 기법이 필요하기 때문에, 높은 예측 정확성과 함께 상대적인 과잉적합 위험성이 적어 최근 많은 연구 대상이 되었던 Support Vector Machine(SVM)을 사용하였다[17, 30]. SVM은 예측의 정확성에도 불구하고 대용량 데이터에의 적용에 한계가 있으므로, 신용카드사 데이터에 적용되기 위해서는 효과적인 데이터 축소 기법과 결합되어야 하며, 정상회원의 대손처리 가능성 예측 문제가 가지고 있는 데이터 불균형 문제도 해결하여야 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 신용카드 회원 대손처리 가능성 문제에 있어서의 예측 대상인 목표 변수를 정의하고 관련 문헌을 정리한다. 제 3장에서는 본 논문에서 사용된 SVM에 대한 이론적 배경을 간단히 정리한 후 SVM과 대

용량 데이터, SVM과 데이터 불균형 문제의 관계에 대해 설명한다. 제 4장에서는 SVM 기반의 대손회원 예측 모형 설계에 대해 설명하면서 데이터 축소 기법으로 2단계 군집화를 제안하고, 데이터 불균형 문제의 해결책으로 앙상블 모형 구성에 대해 설명한다. 제 5장에서는 실험에 사용된 자료 및 실험 결과를 정리한다.

2. 신용카드 회원 대손처리 가능성 예측

2.1 예측 대상의 정의

신용카드 연체회원이 궁극적으로 카드사의 수익성을 악화시키는 것은 장기 악성채권으로 변화되는 경우이다. 단기연체 채권은 채권회수 모형에 의해 관리되고 있으므로 장기악성화, 즉 대손처리가 되지 않는다면 궁극적인 카드사의 피해는 없다. 따라서 본 논문에서는 월 사용액이 상당 금액에 해당하는 신용카드 회원의 대손처리 가능성을 사전 예측하고자 한다.

<표 2>는 국내 한 신용카드사의 2008년 10월~2009년 9월까지의 연체 회원 및 채권을 분석하여 연체 회원의 대손전이 과정을 보인 것으로, 신용카드사에 의해 대환이 이루어질 경우 연체 개월 수가 정지할 수 있음을 감안한 것이다. 통상 신용카드사들은 연체 7개월째부터는 대손처리 회원으로 간주하고 있다. <표 2>에서 볼 수 있듯이 연체 1개월 채권의 경우 상당량이 회수되지만 2개월부터 회수율이 급격히 하락하여 연체 3개월부터는 회수율이 10% 미만에 그쳐 대손 처리될 가능성이 매우 높아진다. 다시 말해 어떤 회원이 신용카드 결재금액을 연체한 상태로 3개월 이상 지속되면 해당 채권은 신용카드사가 회수하기 어려운 채권이 되어 대부분 대손 처리되고 있다.

따라서 본 논문의 예측 대상인 신용카드 대손처리 회원은 연체 3개월 이상인 회원으로 대체하여도 무리가 없을 것으로 판단되며, 예측 시점에서 장래의 불확실한 시점에 대손회원이 되느냐 아니냐를 예측하는 것보다 즉시 연체가 발생하여 3개월 후 3개월 연체 회원이 되는지 여부를 판단하는 것이 문제를 명확히 할 수 있다는 장점이 있다. 이와 같이 신용카드 회원 대손처리 가능성을 예측하는 것은 신용 스코어(Credit Scoring) 모형과 유사한 점이 있다. 신용스코어링 모형을 수립하기 위해서는 회원들을 우량과 불량 회원으로 구분하는데 불량회원의 판정 기준으로 통상 연체 3개월 여부를 많이 사용하고 있기 때문이다. 하지만, 신용스코어링은 일정 시점을 기준으로 대상 회원들을 우량과 불량으로 구분하는데 반해 본 논문에서는 매월 일정 금액 이상 사용한 정상회원들을 대상으로 대손처리 가능성, 즉 예측시점의 다음 달부터 3개월 이상 연체할 가능성이 큰 회원들을 사전 예측한다는 점이 다르다. 이는 특정 월에서 가장 빠

〈표 2〉 연체회원의 대손전이 분석

		당월 연체 개월							
		정상	1개월	2개월	3개월	4개월	5개월	6개월	7개월↑
전월 연체개월	정상	**98%**	**2%**						
	1개월	63%	26%	**11%**					
	2개월	24%	9%	5%	**61%**				
	3개월	5%	2%	1%	2%	**90%**			
	4개월	2%	0%	0%	0%	1%	**93%**		
	5개월	1%	0%	0%	0%	0%	1%	**94%**	
	6개월	1%	0%	0%	0%	0%	0%	0%	**99%**

른 시간 안에 대손 처리되는 회원들을 예측할 수 있도록 해주므로 신용카드사들이 악성채권이 발생하기 전에 대한 선제적 조치를 취할 수 있도록 해주는 장점이 있다. 하지만, 예측대상이 되는 모형의 타깃 회원의 수가 상대적으로 매우 적어지므로 심각한 데이터 불균형 문제가 발생한다.

2.2 관련 문헌

본 논문과 같이 신용카드 대손처리 가능성 예측 문제를 직접 언급한 연구는 찾기 어렵지만, 관련 연구로는 신용 평가 모형(Credit Rating Model) 및 부정행위 적발 모형(Fraud Detection Model)을 들 수 있다. 대상 문제의 응용 분야 관점에서 보면 신용평가 문제에 가깝지만, 문제의 성격을 보면 대손처리 회원의 수가 매우 적기 때문에 심한 데이터 불균형을 보이는 점과 대손 발생 시점을 직접적으로 예측하여야 한다는 측면에서 부정행위 적발 모형과 유사점이 많다.

신용평가나 부정행위 적발 모형 모두 전형적인 이진 분류 문제이며, 로짓 회귀분석, 의사결정수, 신경망 등 전형적인 데이터마이닝 기법이 활용되는 분야이다. 개인과 기업의 신용등급 및 평점 도출, 대출 심사와 개인의 채무 연체 및 기업의 부도 예측 등에 폭넓게 이용되고 있는 신용평가 모형들은 대부분 로짓 회귀분석을 이용하고 있다. 이는 신용평가 모형을 적용할 때 설명 기능이 중요하고, 로짓 모형은 해당 평점의 산출 요인들을 파악하기 용이하기 때문이다[33, 35]. 하지만 최근에 보다 정확한 신용평가를 위한 다양한 데이터마이닝 기법들이 적용되고 있으며[15, 25, 38], 최근에는 SVM과 생존분석, Rough Set 등 다양한 기법을 이용한 연구들이 보고되고 있다[4, 6-9, 23, 24, 27, 28].

예측 대상 변수 즉, 목표변수의 두 범주에 속하는 모집단의 개체 수에 현격한 차이가 있는 경우를 데이터 불균형(Data Imbalance) 문제라고 하는데, 신용카드, 보험사기 및 자금세탁 등의 부정행위 적발시스템, 이동통신 단말기 부정사용, 네트워크 침입방지, 원격경보장치 및 마케팅 분야의 Response Modeling 등의 분야에서 다양하게 나타난다. 이러한 문제들에서는 소수 범주의 정확한 분류가 중요한데 다수 범주의 비율이 높으면 높을수록 정확도 척도가 다수 범주에 의해 왜곡되어지는 현상이 심하게 나타나고, 경우에 따라서는 학습이 전혀 이루어지지 않는 경우도 발생한다[13].

학습 자료의 재구성에 의해 데이터 불균형 문제를 해결하는 방법으로는 첫째, 소수 범주의 수에 맞추어 다수 범주의 데이터를 추출하는 과소추출 방법과 둘째, 다수 범주로부터 추출할 표본크기를 정하고 이에 맞추어 소수 범주로부터 과잉추출 방법이 있다. 또 다른 해결 방법으로는 오분류 비용(Misclassification Cost)을 사용하는 것으로, 소수 범주에 대한 오분류 비용을 다수 범주에 비해 더 크게 적용하는 것이다. 기존의 데이터 분포를 왜곡시키지 않는 장점이 있는 반면에 데이터 불균형이 심할 경우에는 샘플링 방법에 비해 효과가 적다는 단점이 있다[2, 18].

주어진 불균형 데이터 문제에 대한 최적의 학습방법은 사전에 알 수 없지만, Japkowicz and Stephen[26]은 텍스트 분류 문제에 다양한 비율의 과잉, 과소추출에 의한 학습자료를 생성하여 다중모형(Mixture-of-Experts)을 구성한 후, Adaboost 기법을 사용하여 20개의 C4.5로 구성된 모형과 비교하였다. 결과는 다중 모형 즉 앙상블 모형이 학습 및 검증 단계에서 모두 좋은 성능을 보였으며, 특히 검증 단계의 성능이 더 좋았다. 김화경 외 2인[3]은 부정행위 적발 문제에 축소된 앙상블 기법을 사용하였고, 강필성 외 2인[2]의 연구에서는 SVM 앙상블 모형을 사용하여 유사한 결과를 얻었다.

앙상블 학습 방법은 데이터마이닝 분야에서 많은 관심을 받아 왔으며 여러 개의 단위 모형들을 결합하여 예측모형의 정확성 및 안정성을 높여 일반화 능력을 제고시키는 것으로, 첫째 단위 모형들을 어떻게 얼마나 생성할 것인가와 둘째 단위 모형들의 예측치를 어떻게 결합하여 최종 예측치를 얻을 것인가를 해결하여야 한다. 유효한 앙상블 모

형을 얻기 위해서는 단위 모형들의 정확성과 다양성이 확보되어야 하는데 다양성을 얻기 위해 사용되는 방법은 단위 모형별로 적용되는 학습계수, 입력변수 및 학습 자료들을 다르게 적용하는 것이다. 학습자료의 다양성을 추구하는 방법으로 Bootstrapping 샘플을 사용하는 Breiman[11, 12]의 Bagging 기법과 적응적 샘플링을 사용하는 Freund and Schapiro[20]가 개발한 Boosting이 있다. 단위 모형의 적절한 결합에 의해 모형 성능을 개선하려는 노력은 Meta-Learning 또는 Stacked Ensemble이라고 부른다[22, 29, 31, 39].

3. SVM 기반의 분류 모형

3.1 Support Vector Machine

Vapnik[36]에 의하여 제안된 SVM은 고차원 속성 공간(Feature Space)으로 데이터 패턴을 매핑시킴으로써 전역적인 최적화가 가능한 이진 분류 수단으로 주목받았다. 특히, SVM은 미지의 확률분포를 갖는 데이터에 대한 오분류 확률을 최소화하기 위해 구조적 위험 최소화(Structural Risk Minimization) 방법에 근거하고 있어 전통적 기계학습 수단인 인공신경망 등과는 달리 과잉적합 문제로부터 자유롭다는 장점이 있다.

SVM은 기본적으로 선형분리가 가능한 이진분류 문제에서 두 범주 사이의 마진(Margin)을 최대화하는 초평면(Optimal Separating Hyper-plane)을 구함으로써 식 (1)과 같이 표현되는 분류 함수의 일반화 능력을 극대화한다. 여기서 마진은 주어진 초평면으로부터 가장 가까운 패턴까지의 거리의 두 배로 정의되는데 마진을 최대화하는 패턴을 Support Vector(SV)라고 부른다.

$$d(x) = \langle w, x_i \rangle + b = 0 \tag{1}$$

$$Maximize\ J(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}w^T w \tag{2}$$

$$Subject\ to\ \ y_i = (\langle w, x_i \rangle + b) \geq 1,$$
$$i = 1, \cdots, N(Size\ of\ Sample)$$

SVM의 Optimal Hyper-plane은 식 (2)를 최적화함으로써 얻어지는데, Lagrange Multiplier를 적용한 후, KKT(Karush-Kuhn-Tucker) Condition과 Wolfe's Dual Problem을 이용하여 식 (3)과 같이 변형하여 해를 구한다. 식 (3)에서 SVM의 최적 초평면(Optimal Hyper-plane)을 구하는 문제는 한 개의 등식조건과 N개의 부등식 조건을 가진 Quadratic Programming(QP) 문제의 목적 함수를 최대화하는 Lagrange Multiplier를 구하는 것임을 알 수 있다.

$$Maximize\ L(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \tag{3}$$

$$Subject\ to \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\alpha \geq 0,\ i = i, \cdots, N$$

현실의 데이터는 매우 복잡하고 다양하며 잡음을 포함하기 때문에 SVM도 선형 분리가 불가능한 문제에 적용이 가능하도록 확장되어야 한다. 오분류을 인정하는 소프트 마진 기법(Soft Margin Technique)과 주어진 문제를 비선형 SVM으로 확장시키기 위해 커널 트릭(Kernel Trick)을 사용하는 것이 대표적인 SVM의 확장 방법들이다[17].

소프트 마진 기법은 식 (2)의 기본적 SVM에서 고려하지 않았던 오분류를 허용하는 방법으로, 이는 Margin의 폭을 넓혀서 학습하지 않은 패턴들에 대한 일반화 능력을 높이려는 것이다. Soft Margin Technique는 각 패턴마다 오분류에 대한 여유변수(Slack Variable) ξ를 부여하는데, $\xi > 1$이면 오분류된 패턴이며, $0 < \xi < 1$인 경우는 정분류되었지만 마진 내부에 존재하는 패턴이다. 여유변수를 포함하는 SVM 최적화 문제는 식 (4)와 같이 표현된다. C는 오분류된 패턴에 주어지는 페널티 비용(Penalty Cost)이다. 식 (4)의 최적해는 식 (3)에서 $\alpha_i \geq 0$의 조건이 $0 \leq \alpha_i \leq C$으로 변경된 최적화 문제를 풀면 되는데 이는 Lagrange Multiplier가 페널티 비용보다 크지 않다는 조건이 추가된

것이다.

$$Minimize\ J(w)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i \qquad (4)$$
$$Subject\ to\ \ y_i(\langle w, x_i\rangle+b)\geq 1-\xi_i,$$
$$\xi_i\geq 0,\ \ i=1,\cdots,N$$

커널 트릭은 원래의 속성 공간(Feature Space)에서는 선형분리가 불가능하지만, 식 (5)에서와 같이 $Mapping\ Function\ \Phi: x\rightarrow\Phi(x)$을 이용하여 보다 높은 차원의 속성 공간으로 매핑하면 선형분리가 가능하다는 점에 착안한 것이다. 하지만, 고차원으로 매핑시킬수록 계산이 복잡해지고 일반화가 어려워지는 문제점이 있으므로, 실제로 고차원 속성 공간으로 매핑시키지 않으면서도 동일한 효과를 볼 수 있는 커널 트릭을 사용한다. 커널 트릭은 SVM의 최적화 문제에서 패턴 x가 항상 내적(Inner Product) 형태로 나타난다는 점에 착안한 것으로 식 (6)과 같은 성질을 갖는 커널 함수(Kernel Function)를 이용하는데, Mercer's Theorem을 만족하면 Mapping Function ϕ가 존재함이 알려져 있다. 커널 함수를 적용한 SVM 최적화 문제는 식 (7)과 같이 쓸 수 있다.

$$Maximize\ \tilde{L}(\alpha)=\sum_{i=1}^{N}\alpha_i \qquad (5)$$
$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\Phi(x_i)^T\Phi(x_j)$$
$$Subject\ to\ \sum_{i=1}^{N}\alpha_i y_i=0\quad 0\leq\alpha_i\leq C,\quad i=i,\cdots,N$$

$$K(x_1,x_2)=\Phi(x_1)\cdot\Phi(x_2) \qquad (6)$$
$$\tilde{L}(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\Phi(x_i)^T\Phi(x_j) \qquad (7)$$

커널 트릭을 사용하면 명시적으로 Mapping Function을 지정하지 않아도 고차원 속성 공간으로 매핑된 패턴들에 대해 최적화를 수행할 수 있다는 장점이 있는데, Polynomial, Radial Basis Function(RBF), MLP(또는 Sigmoid) 함수 등이 많이 사용되는 커널 함수들이다[32].

3.2 SVM과 대용량 데이터

SVM은 최적 초평면을 얻기 위해 QP 문제에 대한 최적화가 필요한데, 입력 패턴의 차원 및 크기가 증가할수록 계산의 복잡도와 수행시간이 기하급수적으로 증가하는 문제점을 해결하기 위하여 많은 연구가 이루어졌는데 데이터 축소 기법과 최적화 알고리즘 개선의 두 가지로 구분할 수 있다.

첫 번째의 데이터 축소 방법은 주어진 학습자료를 대표할 수 있는 대표적인 패턴들만 선택하여 SVM의 학습에 사용하는 방법으로, 단순 임의추출 방법부터 Hierarchical Clustering, Parallel Clustering 및 K-means Clustering 등의 다양한 군집화 방법이 사용되었다[10, 42]. 최근 Cervantes et al.은 Core-Set 및 Enclosing Ball Clustering을 이용하여 효과적인 SVM 학습자료 구성을 보여 주었다[14].

두 번째는, SVM 최적화 알고리즘을 대용량 데이터 처리가 가능하도록 수정하는 방법으로 기본적으로 Subset Selection Method(SSM)에 기반을 두고 있다. SVM에서는 학습자료 중에서 특정 패턴만이 Lagrange 승수의 조건을 만족하여 SV가 되므로, SV 집합을 알 수 있다면 이들만으로 학습자료를 구성하여 최적 초평면을 도출하는 것이 가능하다. 하지만 SV 집합을 사전에 알 수 없기 때문에 SSM은 학습자료의 일부를 임의로 SV 집합으로 가정한 후 최적화 작업을 반복 수행하여 최적의 초평면을 구하는 것이다. Projected Conjugate Gradient(PCG) Chunking 알고리즘과 Sequential Minimum Optimization(SMO) 알고리즘이 대표적이다. 특히 SMO는 SV 집합의 크기를 2로 가정하여 반복 수행함으로써 QP 문제의 크기를 가능한 가장 작은 크기의 하위 QP 문제로 분해하기 때문에 PCG Chunking보다 학습 속도가 빠른 것으로 알려져 있다[16, 19, 37].

3.3 SVM과 데이터 불균형 문제

SVM도 데이터 불균형으로 인한 예측성능의 저하를 피할 수 없는데, 특히 과도하게 분포된 다수 범주에 의해 의사결정 경계가 모호해지므로 Soft-Margin Technique을 적용하는 경우, Cost Factor가 매우 크게 설정하지 않으면 대부분의 학습 패턴을 다수 범주로 분류하는 현상이 발생한다. 이는 SVM이 총 오분류 비용을 최소화하려 하기 때문이다. 이에 대한 해결 방안의 첫 번째는 SVM의 최적 초평면을 도출하는 알고리즘을 수정하는 것으로 다수 범주와 소수 범주에 각기 다른 Cost Factor를 적용하거나 Kernel Function 및 목적함수를 수정하는 방법이다[40, 41]. 두 번째는 앞서 설명한 바와 같이 표본에 대한 과소추출 또는 과잉추출에 의한 학습자료의 재구성 방법을 적용하는 것이며, 마지막으로 여러 개의 SVM 모형을 생성하여 앙상블 모형을 구성하는 것이다[5, 21, 34].

4. 대손처리 가능성 예측 모형의 설계

4.1 SVM 기반 분류 모형의 설계

본 논문의 목적은 신용카드 대손처리 회원을 사전 예측하여 신용카드사의 수익성 악화를 방지하기 위하여 매월 일정 금액 이상 사용한 정상회원을 대상으로 결제일 직후 대손처리 가능성이 큰 회원을 선별하는 것이다. 이를 위해 예측기준 시점 다음 달부터 연체 상태를 3개월 동안 계속 유지할 가능성이 큰 회원들을 분류해내는 SVM 기반 이진 분류 모형을 설계하였는데, 입력 변수로는 회원의 기본 신상정보, 신용정보 및 최근 6개월 동안의 신용카드 사용행태에 관한 가공정보 등이 사용되었는데 <표 3>에 요약하였다. SVM 기반 신용카드 정상회원 대손처리 가능성 예측 모형을 개발하기 위해서는 대용량 데이터를 SVM 학습이 가능하도록 효과적인 데이터 축소 기법의 개발과 함께 대손처리 회원의 비율이 매우 낮다는 사실 즉 데이터 불균형 문제를 해결하여야 한다.

4.2 2단계 군집화에 의한 데이터 축소

월별 대손가능성 예측을 위하여 월 단위로 학습자료를 구성하여야 하는 점과 신용카드사의 회원규모를 고려하면 효과적인 SVM 학습을 위한 데이터 축소 과정이 반드시 필요하다. 본 논문에서 사용한 데이터 축소 방법은 개인의 소비 성향 및 수입/지출의 규모는 단기간에 크게 변동하지 않는다는 사실에 착안하여 개발된 2단계 군집화 과정으로 SV가 될 수 있는 가능성이 높은 학습 패턴들을 추출하는 것이다.

첫 번째 단계의 군집화는 동일한 회원의 데이터를 축소시키는 것이다. 대손처리 가능성을 예측하는 것이므로 연체가 없는 정상회원들이 압도적으로 우위에 있는 다수 범주에 속하게 되므로, 1차적으로 다수 범주에 해당하는 패턴들을 줄이려는 것이다. 신용카드를 이용하는 회원의 결제일은 매월 도래하므로 일정 금액 이상 사용한 모든 회원에 대하여 월 1회씩 대손처리 가능성을 예측하기 위해서는 [그림 1]과 같이 입력 자료들을 생성하여야 한다. [그림 1]에서 볼 수 있듯이 입력 패턴의 생성 시 5개월의 집계기간 중복이 발생하므로 만약, 해당 회원의 소비 성향 및 수입/지출의 규모에 큰 변화가 없었다면 해당 회원의 월별로 생성된 입력패턴들은 다른 회원들의 입력패턴과 비교하여 유사도가 매우 높을 것으로 판단된다. 다시 말해 정상회원의 경우, 개인의 카드 사용행태나 신용 관련 사항 등에 큰 변동이 없었다면 동일 회원의 월별 패턴 사이의 유사성은 매우 높을 것이다.

이와 같이 입력 패턴의 생성에 사용되는 정보들 중 특히 장기간에 걸쳐 집계, 가공된 정보는 예측시점과 무관하게 유사한 값을 보일 가능성이 높다. 따라서 논문에서는 우선 다수 범주를 차지하는 정상 회원들에 대하여 각 회원별로 해당 입력패턴들만 가지고 정규 화된 유클리디언 거리를 이용하여

군집화 작업을 수행하였다. 이 과정에서 입력패턴의 유사도 차이 정도에 따라, 단일 회원의 입력패턴은 최소 1개에서 최대 생성된 입력패턴 전부까지 변동될 수 있다.

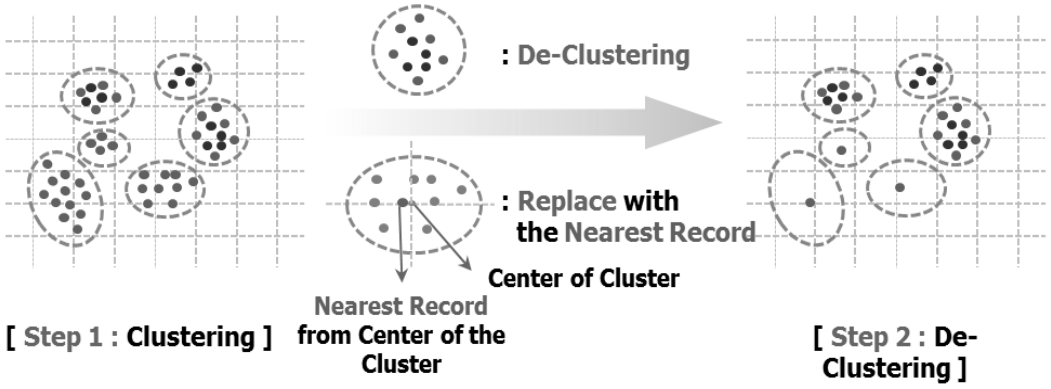
〈표 3〉 입력변수로 사용된 정보들

원천 데이터	사용된 입력변수들의 예
신용카드 및 회원 기본정보	성별, 나이, 거주지, 직업, 신용카드 개설일 등
신용카드 승인 내역	거래 및 승인유형, 승인금액, 할부거래 여부 등
결재대금 입금 내역	입금일, 입금 금액 등
타사 신용카드 이용 내역	총 한도금액, 신판 이용금액, 현금서비스 이용금액, 연체금액 등
Credit Bureau 제공 정보	회원 신용등급 및 평점 등
여신 관련 정보	신용대출 이용횟수와 금액, 회원의 이용 기관 수, 담보금액 등
연체 관련 정보	연체 시작 및 종료일자, 연체 금액, 집계시점의 연체 여부 등

두 번째 단계의 군집화는 정상회원의 입력패턴들 중에서 상대적으로 SV가 될 확률이 낮은 패턴들을 제거하는 과정으로, SVM의 적용을 용이하게 해주는 데이터 축소 작업을 하면서 데이터 불균형 문제에 대한 고려를 함께 반영한 것이다. 효과적인 학습자료의 구성을 위해서는 상대적으로 희소하고 예측 모형의 성능에 중요한 영향을 미치는 정보는 보존하고 그렇지 못한 정보들은 가능한 제거하는 것이 바람직하다. 따라서 희소한 대손처리 회원의 입력패턴들과 상대적으로 SV 집합에 속할 확률이 높은 정상회원의 입력패턴들만 학습자료에 보존하려는 것이다.

상대적으로 SV 집합에 속할 확률이 높은 패턴이란 SVM의 초평면과 가까운 거리에 있는 패턴들, 즉 대손회원과 정상회원의 입력패턴이 함께 분포하는 군집인 경우이며, 데이터의 대부분을 점유하는 정상회원의 입력패턴들만 분포하는 군집인 경우는 해당 패턴들이 SV가 될 확률이 매우 낮다. 따

[그림 1] 6개월 단위로 집계 생성 입력 변수



[그림 2] 2단계 군집화의 두 번째 군집화 과정

라서 정상회원들만 분포하는 군집이 데이터 축소의 대상이 되는데, 데이터 축소 과정을 도식화한 것이 [그림 2]이다.

두 번째 단계의 군집화는 [그림 2]의 좌측과 같이 첫 번째 군집화의 결과, 즉 회원별 입력패턴의 제거 결과 남은 입력패턴 전체에 대해 군집화를 적용하여 대손회원만 분포하는 군집, 정상회원만 분포하는 군집 및 두 집단이 동시에 분포하는 군집으로 선별한 후 데이터 축소 작업을 하는 것이다. 대손회원이 포함된 군집들은 그대로 보존하고, 정상회원들의 입력패턴만으로 분포된 군집들은 군집의 중심에서 유클리디언 거리가 가장 가까운 패턴만 남기고 나머지는 모두 제거한다. [그림 2]의 좌측과 우측, 즉 군집화 수행 전과 후 모두, 2차원 평면의 우측 상단에 SVM의 최적 초평면이 위치할 것임을 가시적으로 예상할 수 있다. 이와 같이 SVM의 초평면 결정에 영향을 미치지 않으면서 데이터 축소가 가능하다는 것과 데이터 불균형 문제도 어느 정도 해소하여 Soft Margin Technique의 오분류(Miss-Classification) 비용이 다수 범주 위주로 모형을 생성하는 것을 방지하는 기능을 유지할 수 있다는 것이다.

본 논문에서는 2단계 군집화에 의한 데이터 축소 기법을 적용하면서 유효성을 검증하기 위하여 SVM의 학습자료를 단계별로 세분화하고, 임의추출에 의한 학습자료도 함께 사용하여 검증하였다. 본 논문에서 사용된 학습자료들을 정리하면 <표 4>와 같으며, 대손처리 회원으로 판정된 회원들의 입력패턴, 즉 본 논문에서 소수범주에 해당하는 목표 범주(Target Class)에 속하는 입력패턴들은 모든 학습자료에 동일하게 포함시켰다.

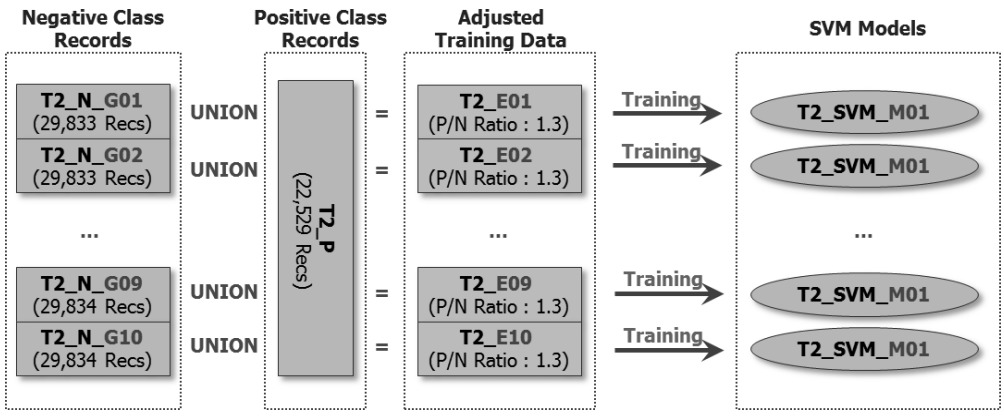
〈표 4〉 학습자료의 구분

학습자료 구분	설명
T0	데이터 축소 전의 학습자료
T1	정상회원의 1단계 군집화 후의 학습자료
T2	T1에 대해 2단계 군집화 후의 학습자료
S1	T0으로부터의 정상회원 임의 추출
S2	T1으로부터의 정상회원 임의 추출

4.3 SVM 앙상블 모형

본 논문의 신용카드 회원 대손처리 가능성 예측 문제는 2단계 군집화에 의한 데이터 축소 후에도 여전히 데이터 불균형 문제가 존재하므로 이에 대한 해결책으로 학습자료의 다양화에 의한 SVM 앙상블 모형을 [그림 3]과 같이 구성하였다.

학습자료 T2를 사용하여 10개의 SVM 모형을 생성한 후 앙상블 모형을 구성하였다. 각 단위 SVM 모형은 T2로부터 임의추출에 의해 10개의 하위



[그림 3] SVM 앙상블 모형의 구성

학습자료, 즉 정상회원들의 입력패턴을 추출한 후, 대손회원들의 입력패턴과 합하여 단위 SVM 모형을 위한 학습자료들을 구성하였다. 하위 학습자료들을 생성하는 과정에서 정상/대손회원 입력패턴의 균형을 맞추기 위하여 과잉추출 기법을 적용하여 데이터 불균형으로 인한 학습의 어려움이 없도록 하였다. SVM의 성격을 고려하여 최종 앙상블 모형의 결과는 Voting을 사용하여 도출하였다.

5. 실험

5.1 데이터 수집

본 논문에 사용된 데이터는 국내 비은행계 신용카드사의 2008년 1월부터 2009년 11월 사이에 실제로 신용카드 사용이력이 있는 회원들을 대상으로 수집되었다. 하지만 입력 자료의 생성에 과거 6개월 동안의 이용실적을 집계 사용하는 점과 본 연구의 목표변수인 대손처리 여부를 결정하기 위하여 연체 3개월 여부를 확인하여야 하므로 실제 학습자료에는 2008년 7월부터 2009년 8월 사이에 월 이용대금이 500,000원 이상인 정상회원으로 한정하였다. 입력패턴의 생성 시에는 정상이지만 실제로 3개월 후에 대손 회원으로 확인된 회원들은 선별하여 목표 범주에 포함시켰다. 수집된 학습자료 중 2009년 7, 8월 자료는 학습된 모형의 검증을 위한 테스트 집합으로 사용했으며 나머지 자료를 SVM의 훈련 집합(Training Set)으로 사용하였다. 훈련 집합의 경우 약간의 데이터 전처리 과정을 거쳤는데 우선 월 중 결제일이 변경된 회원들의 경우 변경된 결제일의 첫 번째 패턴은 중복되므로 이를 삭제하였다. 또, 연체 3개월인 회원의 연체 시작 전 2개월, 3개월째 생성된 패턴은 정상으로 파악됨에도 집계변수들의 생성 과정을 살펴볼 때, 연체 3개월이 예상되는 월, 즉 목표변수인 대손처리 회원으로 판정한 월의 입력패턴과 큰 차이를 보이지 않을 가능성이 높다. 이러한 현상은 SVM의 의사결정 경계를 모호하게 만들 가능성이 높으므로 훈련 집합에서 제거하였다.

SVM 학습에 있어 대용량 데이터 문제를 해결하기 위하여 2단계 군집화 과정을 도입하였는데, <표 5>와 같은 11개의 변수들을 유클리디언 거리 기반의 K-Means 군집화 기법에 사용하였다. 특히, 두 번째 군집화 과정에서 초기 군집 수를 지정 문제가 있었다. 본 실험에서는 5,000, 10,000, 15,000으로 변화시키며 실험을 수행한 결과, De-Clustering 후 정상 범주의 레코드 수가 610,209건, 298,334건, 331,789건으로 축소되었는데 초기 군집의 수가 10,000일 경우를 선택하였다. <표 6>는 이상의 과정을 거쳐서 생성된 학습자료를 요약한 것이다.

〈표 5〉 2단계 군집화에 사용된 변수들

범주	변수명
회원 기본정보	Platinum 카드 여부
신용카드 이용패턴	당월 신판 이용금액 당월 할부 이용금액 당월 현금서비스 이용금액 다빈도 금액 범주
신용정보	Credit Bureau 신용등급 B
연체 및 입금률	최근 6개월 연체횟수 최근 6개월 최대 연체금액
이용기관 및 한도정보	총이용기관수 리볼빙이용기관수
신용대출	최근 1년 신용대출건수

〈표 6〉 실험에 사용된 학습자료

학습자료의 구분		범주의 구성		
		정상	대손	정상 : 대손
훈련 집합	T0	10,896260	22,529	483 : 1
	T1	2,386,671	22,529	106 : 1
	T2	298,334	22,529	13 : 1
	S1	2,386,671	22,529	106 : 1
	S2	298,334	22,529	13 : 1
테스트 집합	2009. 7	1,044,888	1,288	811 : 1
	2009. 8	1,022,037	1,191	858 : 1

5.2 SVM의 학습

본 논문에서는 SVM 학습을 위하여 커널 함수로

$k(x_i, x_j)=\exp(-\gamma \| x_i - x_j \|^2)$과 같이 표현되는 RBF 함수를 사용하였고, 최적화 방법으로 SMO를 사용하였다. SVM의 학습 파라미터인 Cost Factor C와 RBF 커널 파라미터인 γ를 결정하기 위하여 5-Fold Cross Validation을 적용한 Grid Search를 사용하였다. Grid Search는 많은 시간이 소요되기 때문에 비교적 데이터 크기가 작은 학습자료 'T2'를 사용하였고, 최종 결정된 값을 나머지 SVM 모형 생성에 동일하게 적용하였다. 최적 파라미터를 결정함에 있어 통상의 정확도(Total Accuracy) 척도는 데이터 불균형으로 인하여 정확한 판단을 어렵게 하므로 True Positive와 False Positive의 비율을 Target Accuracy로 정의하여 사용하였다. <표 7>에서 각 셀의 수자들은 True Positive, False Positive 및 두 수의 비율을 나타낸 것이다. Cost Factor는 2^4, RBF 커널 파라미터는 2^{-4} 경우가 True Positive 대비 False Positive 비율이 2.44로 가장 낮아서 SVM의 파라메타로 결정하였다.

SVM 앙상블 모형은 학습자료 'T2'에만 적용하였다. 앙상블 모형이 데이터 불균형 문제에 대한 해결책으로 사용하는 것이므로 2단계 군집화에 의한 데이터 축소 기법의 타당성 검증을 위한 학습자료들인 T1, S1, S2에 적용할 필요는 없다고 판단하였기 때문이다.

5.3 실험결과

2단계 군집화에 의한 데이터 축소가 SVM의 학습에 미치는 영향을 파악하기 위하여 <표 7>에 각 훈련 집합에 대한 SVM의 학습 결과를 정리하였다. SVM의 학습 결과를 두 가지 정확도 측면에서 비교해보면 T2 Ensemble, T2, T1, S2, S1 순서로 나타났다. 특히, 'T1'을 임의 추출하여 얻어진 'S2'를 훈련 자료로 이용한 모형이 'T1'을 이용한 모형보다 학습 성능이 떨어진다는 것은 2단계 군집화 기법이 유효했음을 입증하는 것으로, SVM 학습에 있어 SV가 될 가능성이 높은 입력패턴의 선별 중요성을 보여준 것이다. 데이터 불균형을 감안하면 예상대로 'T2 Ensemble', 즉 T2 학습자료에 SVM 앙상블을 적용한 경우의 학습 결과가 가장 좋았고, 학습자료 크기 감소로 인하여 학습에 소요된 CPU Time도 가장 적었다. 앙상블 모형

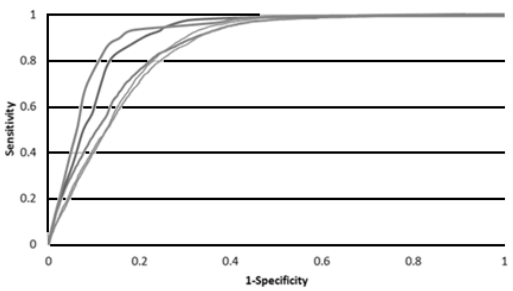
〈표 7〉 Grid Search에 의한 SVM 파라미터 결정

c \ γ	2^5	2^4	2^2	2^{-2}	2^{-4}	2^{-5}
2^{-5}	2,068 7,094 **3.43**	2,156 7,129 **3.31**	2,326 7,241 **3.11**	2,562 7,345 **2.87**	2,985 7,945 **2.66**	3,054 7,959 **2.61**
2^{-4}	2,420 7.621 **3.15**	2,358 7,803 **3.31**	2,508 7,903 **3.15**	2,489 7,729 **3.11**	2,890 7,920 **2.74**	2,993 8,003 **2.67**
2^{-2}	2.576 7.689 **2.98**	2,491 7,830 **3.14**	2,451 7,513 **3.07**	2,942 8,024 **2.73**	2,972 8,120 **2.73**	2,891 8,940 **3.09**
2^2	2,995 8.342 **2.79**	2,943 8,103 **2.75**	3,018 8,120 **2.69**	3,094 7,843 **2.53**	3,241 8,234 **2.54**	3,480 8,620 **2.48**
2^4	3.208 8.125 **2.53**	3,054 8,120 **2.66**	3,189 8,106 **2.54**	3,099 7,931 **2.56**	3,481 8,510 **2.44**	3,109 8,421 **2.71**
2^5	3,157 8,295 **2.63**	3,304 8,690 **2.63**	3,208 8,105 **2.53**	3,089 7,918 **2.56**	3,309 8,096 **2.45**	3,209 8,390 **2.61**

〈표 8〉 훈련 자료에 따른 SVM의 학습 결과

		Training Data Set				
		T1	T2	S1	S2	T2 Ensemble
Actual Positive	True Positive	2,048	2,096	2,045	2,041	2,088
	False Negative	431	383	434	438	391
Actual Negative	True Negative	1,579,459	1,835,917	1,528,841	1,535,891	1,910,904
	False Positive	487,466	231,008	538,084	531,034	156,021
Sensitivity		0.8261	0.8455	0.8249	0.8233	0.8423
Specificity		0.7613	0.8869	0.7397	0.7431	0.9139
Target Accuracy		0.0042	0.0090	0.0038	0.0038	0.0132
Total Accuracy		0.7614	0.8868	0.7398	0.7432	0.9244
Training CPU Times (HH : MM : SS)		23 : 29 : 11	10 : 45 : 03	24 : 01 : 32	10 : 23 : 49	05 : 31 : 53

임에도 학습에 소요된 CPU 시간이 가장 적었다는 것은 SVM의 최적화에 소요되는 시간이 데이터 크기에 매우 민감함을 보여주는 것이다. <표 8>의 학습 결과를 보다 명확히 하기 위해 [그림 4]에서와 같이 ROC Curve를 작성하였는데 예상대로 'T2 Ensemble'과 'T2'가 가장 좌측 상단에 위치하였다.



[그림 4] 훈련 집합에 따른 ROC Curves

학습된 SVM 모형들의 예측 성능을 평가하기 위하여 2009년 7, 8월 데이터를 테스트 집합으로 사용하였다. 본 논문이 제안한 2단계 군집화에 의한 SVM 학습 방법의 예측 성능을 객관적으로 평가하기 위하여 로짓 회귀분석(Logit Regression, LT), 의사결정수(Decision Tree, TR) 및 신경망(Artificial Neural Network, ANN)을 사용하였다. 성능 비교를 위한 세 기법 모두 동일한 입력변수를 사용하였으며, 의사결정수와 신경망의 경우 학습 과정에서 검증 집합(Validation Set)이 필요하므로 각 훈련 집합을 7 : 3으로 나누어 사용하였다. 학습에 사용된 신경망은 2개의 은닉층에 각각 12개와 6개의 노드를 갖는 구조이며, 최대 학습 횟수는 250회로 제한하였다. <표 9>, <표 10>는 테스트 집합에 대한 각 기법의 예측 성능을 정리한 것이다. 예를 들어 <표 9>에서 ANN-T2는 신경망을 훈련 집합 T2에 학습시킨 후 테스트 집합에 대하여 성능 평가를 한 것이다. 평가의 공정성을 위하여 모든 학습의 경우에 대하여 True Positive의 수를 일정 수준에 맞춘 후 평가를 하였으며, 평가 결과를 보완하기 위하여 <표 10>의 누적 Lift 도표를 작성하였다.

<표 10>에서 볼 수 있듯이 SVM-T2, ANN-T2, LT-T2, DT-T2 등의 순위로 예측 성능이 좋았다. 특히, 훈련 집합에 비하여 테스트 집합의 목표 범주의 비율이 484 : 1에서 834 : 1 높아져 정확한 분류에 의한 대손회원 예측이 어려워졌음에도 'T2' 훈련 집합에 학습된 SVM과 신경망이 좋은 예측 성능을 보인 것은 2단계 군집화에 의한 데이터 축소가 효과적으로 SV를 선별 유지하여 의사결정

〈표 9〉 SVM 예측 성능의 평가

		SVM-T1	SVM-T2	LT-T1	LT-T2	DT-T1	DT-T2	ANN-T1	ANN-T2
Actual Positive	True Positive	2,048	2,096	2,037	2,041	2,021	2,039	2,014	2,056
	False Negative	431	383	442	438	458	440	465	423
Actual Negative	True Negative	1,579,459	1,835,917	1,561,456	1,751,464	1,452,412	1,632,346	1,568,174	1,772,079
	False Positive	487,466	231,008	505,469	315,461	614,513	434,579	498,751	294,846
Sensitivity		0.8261	0.8455	0.8217	0.8233	0.8152	0.8225	0.8124	0.8294
Specificity		0.7613	0.8869	0.7554	0.8474	0.7027	0.7897	0.7587	0.8574
Target Accuracy		0.0042	0.0090	0.0040	0.0065	0.0033	0.0047	0.0040	0.0070
Total Accuracy		0.7614	0.8868	0.7555	0.8473	0.7028	0.7898	0.7588	0.8573

〈표 10〉 누적 LIFT에 의한 예측 성능의 평가

Decile	Cumulative Lift							
	SVM-T1	SVM-T2	LT-T1	LT-T2	DT-T1	DT-T2	ANN-T1	ANN-T2
1	5.1150	5.8814	5.0150	5.1953	3.4490	4.0031	5.0126	5.3731
2	3.6487	4.3687	4.0393	4.2194	3.3481	3.3249	3.4174	4.2959
3	2.9071	3.1532	3.1284	3.0154	2.8546	2.8451	2.7623	2.9488
4	2.3659	2.4355	2.4459	2.1546	2.3689	2.3246	2.2486	2.3760
5	1.9581	1.9758	1.9812	1.9632	1.9589	1.1954	1.9476	1.9589
6	1.6411	1.6559	1.6585	1.6497	1.6478	1.6324	1.6484	1.6492
7	1.4142	1.4234	1.4254	1.4281	1.4165	1.4246	1.4985	1.4188
8	1.2414	1.2470	1.2483	1.2489	1.2450	1.2132	1.2345	1.2455
9	1.1071	1.1093	1.1106	1.1042	1.1089	1.1034	1.1241	1.1084
10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

경계를 단순화함으로써 신용카드 대손회원 예측에 매우 유용한 방법임을 보여 주는 것이다. 또, <표 9>의 누적 Lift Chart에서도 볼 수 있듯이 2단계 군집화를 이용한 SVM의 예측 성능이 전체적으로 가장 좋은 성능을 보였다.

이상의 실험 결과로 이진 분류 문제에 있어 SVM은 신경망 등과 같은 기존의 분류 기법보다 강력하다는 것을 알 수 있다. 하지만, SVM이 현실에서 사용되기 위해서는 로짓 회귀분석과 같이 목표 클래스에 속할 확률 값을 제공할 수 있어야 한다. 이를 위해 SVM에서는 각 패턴으로부터 분류의 경계면을 형성하는 초평면까지의 정규화된 거리(Normalized Distance)를 사용한다. 다시 말해 SVM에서는 목표 범주로 분류된 패턴은 양의 거리 값을 가지고 반대의 경우는 음의 거리 값을 가지게 되므로 Min-Max Normalization을 이용하여 SVM의 출력값을 0과 1사이의 값으로 정규화하면 확률의 대체 값으로 사용할 수 있다. <표 11>은 SVM의 출력값을 확률로 변형하여 확률 기준값(Probability Threshold) 따른 예측 성능의 변화를 보인 것으로, 이 경우에도 2단계 군집화를 사용한 훈련 집합 T2에 학습시킨 SVM의 예측 성능이 더 우수함을 알 수 있다.

6. 결 론

신용카드 연체회원 예측 모형의 하나로 대손처리 가능성을 예측하는 모형을 SVM을 적용하여 개발하였다. 본 논문에서는 예측 대상인 대손 회원을 연체 3개월째 회원으로 대체하여 연체 시작 전에 즉, 아직 정상회원인 상태에서 회원의 결제일에 대손처리 가능성을 매월 예측하였다. SVM

〈표 11〉 확률 기준값에 따른 SVM의 예측 성능

Training DataSet	Probability Threshold	True Positive	False Negative	True Negative	False Positive	Sensitivity	Specificity	Accuracy
T1	0.9993	4	2,475	2,066,546	379	0.0016	0.9998	0.9986
	0.9984	31	2,448	2,065,199	1,726	0.0125	0.9992	0.9980
	0.9971	62	2,417	2,062,329	4,596	0.0250	0.9978	0.9966
				…				
	0.5775	1,944	535	1,630,585	436,340	0.7842	0.7889	0.7889
	0.5586	1,969	510	1,618,347	448,578	0.7943	0.7830	0.7830
	0.5418	1,992	487	1,607,231	459,694	0.8035	0.7776	0.7776
				…				
	0.0408	2,438	41	986,835	1,080,090	0.9835	0.4774	0.4780
	0.0133	2,470	9	589,988	1,476,937	0.9964	0.2854	0.2863
	0	2,479	0	0	2,066,925	1.0000	0.0000	0.0012
T2	0.9998	2	2,477	2,066,706	219	0.0008	0.9999	0.9987
	0.9985	40	2,439	2,065,828	1,097	0.0161	0.9995	0.9983
	0.9981	72	2,407	2,064,438	2,487	0.0290	0.9988	0.9976
				…				
	0.5335	2,005	474	1,832,410	234,515	0.8088	0.8865	0.8864
	0.5109	2,035	444	1,833,384	233,541	0.8209	0.8870	0.8869
	0.5001	2,101	378	1,835,906	231,019	0.8475	0.8882	0.8882
				…				
	0.3257	2,449	30	1,121,769	945,156	0.9879	0.5427	0.5433
	0.0259	2,477	2	582,304	1,484,621	0.9992	0.2817	0.2826
	0	2,479	0	0	2,066,925	1.0000	0.0000	0.0012

을 이용하여 신용카드 대손 회원을 예측하기 위해서는 대용량 데이터의 축소 문제와 데이터 불균형 문제를 극복하여야 한다. 본 논문에서는 2단계 군집화 기법을 사용하여 SVM의 Support Vector가 될 가능성이 높은 입력패턴을 선별하면서 데이터를 축소시키는 방안을 제시하였으며, 데이터 불균형 문제를 해결하기 위하여 SVM 앙상블 모형을 사용하였다.

국내 신용카드사의 자료를 이용하여 검증한 결과 2단계 군집화에 의해 생성된 학습자료에 앙상블 기법을 적용한 SVM 모형이 가장 우수한 예측성능을 보였으며, 개발된 모형은 로짓 회귀분석, 의사결정수, 신경망 등과 비교하였을 때도 가장 우수한 예측 성능을 보였다. 따라서 SVM은 적절한 데이터 축소 기법과 함께 사용될 수만 있다면 현실의 여러 분류 문제에 적용될 수 있음을 보였다. 신용카드사의 대손 처리는 경기 침체기에 대량으로 발생되므로 사전에 이를 예측하여 대비할 수 있다면 신용카드사의 수익성 개선은 물론 경기 침체가 예상되는 시점에서 불량가능성이 높은 회원들에 대한 조치가 가능해지는 장점이 있다.

본 논문은 대손회원을 예측함에 있어 예측시점에서 정상회원을 대상으로 익월부터 3개월간 연체상태를 유지될 회원을 예측하기 때문에 심각한 데이터 불균형 문제가 발생한다. 따라서 문제 해결 과정은 신용평가 모형보다는 부정행위 적발 모형에 더 가깝고 예측 모형의 False Positive Rate에 매우 높게 나타날 가능성이 있음을 의미한다. 이러한 어려움을 극복하는 방안은 다양한 변수 가공을 통한 유효한 입력변수의 개발 및 적합한 앙상블 모형의 구성을 발견하는 것이다. 특히, SVM의 장점에도 불구하고 데이터 불균형이 심한 문제에서

는 앙상블 모형의 유효성이 입증되었으므로 이를 발전시키기 위한 추가 연구가 필요하다. 또, 본 논문에서는 2008년 글로벌 금융위기에 따른 신용위험이 높아진 시기의 자료를 사용하였는데 평상시의 자료를 수집하여 본 연구의 타당성을 폭넓게 검토할 필요도 있다. 이상과 같은 추가 연구가 성공적으로 금융기관은 거시경제지표에 의존한 단순한 조기경보시스템(EWS, Early Warning System)이 아니라, 구체적인 미시적 조치가 가능한 효과적인 조기경보시스템의 구축이 가능해 질 것이다.

참 고 문 헌

[1] 금융감독원 보도자료, 「신용카드사 경영실적」, 2002~2006.

[2] 강필성, 이형주, 조성준, "데이터 불균형 문제에서의 SVM 앙상블 기법의 적용", 「한국정보과학회 추계학술대회논문집」, 제31권, 제2호(2005), pp.706-708.

[3] 김화경, 한상범, 지원철, "축소된 앙상블을 이용한 현금융통 적발 모형", 「지능정보연구」, 제16권(2010), pp.93-116.

[4] 노태협, 유명환, 한인구, "러프집합 이론과 사례기반추론을 결합한 기업신용평가 모형", 「정보시스템연구」, 제14권(2005), pp.41-65.

[5] 이영섭, 오현정, 김미경, "데이터마이닝에서 배깅, 부스팅, SVM 분류 알고리즘 비교 분석", 「응용통계연구」, 제18권(2005), pp.343-354.

[6] 이영찬, "인공신경망과 Support Vector Machine의 기업부도예측 성과 비교 : Support Vector Machine의 유용성을 중심으로", 「한국지능정보시스템학회 2004년 춘계학술대회 논문집」, 2004.

[7] 정석훈, 서영무, "Rough Set 기법을 이용한 신용카드 연체자 분류", 「Entrue Journal of Information Technology」, 제7권(2008), pp.141-150.

[8] 하성호, 양정원, 민지홍, "코호넨 네트워크와 생존분석을 활용한 신용예측", 「한국경영과학회지」, 제34권(2009), pp.35-54.

[9] Allen, L. N. and L. C. Rose, "Financial survival Analysis of default debtors", *Journal of the Operational Research society*, Vol.57(2006), pp.630-636.

[10] Awad, M., L. Khan, F. Bastani, and I. L. Yen, "An effective support vector machine SVMs performance using hierarchical clustering", *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence(ICTAI)*, 2004.

[11] Breiman, L., "Bagging Predictors", *Machine Learning*, Vol.24(1996), pp.123-140.

[12] Breiman, L., "Arcing Classifiers", *Annals of Statistics*, Vol.26(1998), pp.801-849.

[13] Chawla, N. V., N. Japkowicz, and A. Kolcz, "Editorial : Special Issue on Learning from Imbalanced Data Sets", *SIGKDD Exploration*, Vol.6(2004), pp.1-6.

[14] Cervantes, J., X Li, and W Yu, "Support vector machine classification for large data sets via minimum enclosing ball clustering", *Neurocomputing*, 2008.

[15] Chen, M. C. and Huang, S. H., "Credit Scoring and Rejected Instances Reassigning through Evolutionary Computation Techniques", Expert Systems with Application, Vol.24(2003), pp.433-441.

[16] Collobert, R. and S. Bengio, "SVMTorch : Support vector machines for large regression problems", 2001.

[17] Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[18] Dietterich, T., "An Experimental Comparison of Three Methods for Constructing En-

sembles of Decision Trees : Bagging, Boosting and Randomization", *Machine Learning*, Vol.40, No.2(2000), pp.139-157.

[19] Fan, R. E. and P. H. Chen, "Working set selection using second order information for training SVM", *Journal of Machine Learning Research*, 2005.

[20] Freund, Y. and R. Shapiro, "A Decision-theoretic Generalization of On-line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, Vol.55 (1997), pp.119-139.

[21] Gustavo, E. A., P. A. Batista, and R. C. Prati, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data", *SIGKDD Explorations*, 2004.

[22] Hansen, L. and P. Salomon, "Neural Network Ensembles", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.12(1990), pp.993-1001.

[23] Huang, Z., H. Chen, C. J. Hsu, and W. H. Chen. "Credit rating analysis with support vector machines and neural networks : a market comparative study", *Decision support systems*, 2004.

[24] Huang, C. L. and M. C. Chen. "Credit scoring with a data mining approach based on support vector machines", *Expert Systems with Applications*, 2007.

[25] Hsigh, N. C., "Hybrid mining approach in the design of credit scoring models", *Expert Systems wih Application*, Vol.28(2005), pp.655-665.

[26] Japkowicz N. and S. Stephen, "The Class Imbalance Problem : A Systematic Study", *Intelligent Data Analysis*, Vol.6, No.5(2002), pp.429-450.

[27] Min, J. H., C. W. Jeong, and M. S. Kim, "Tuning the Architecture of Support Vector Machine : The Case of Bankruptcy Prediction", *Int'l Journal of Management Science*, Vol.17, No.1(2011), pp.19-43.

[28] Min, J. H,. "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters", *Expert Systems with Applications*, 2005.

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

[30] Platt, J., *Advances in Kernel Methods : Support Vector Machine : Fast training of support vector machine using sequential minimal optimization*, MIT Press, 1998.

[31] Rooney, N., D. Patterson and C. Nugent, "Pruning Extension to Stacking", *Intelligent Data Analysis*, Vol.10(2006), pp.47-66.

[32] Scholkopf, B., K. K. Sung and C. J. C. Burges, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers", *Signal Processing*, 2002.

[33] Siddiqi, N., *Credit Risk Scorecards*. John Wiley and Sons, 2006.

[34] Tang, Y., Y. Q. Zhang, and N. V. Chawla, "SVMs Modeling for Highly Imbalanced Classification", *IEEE Transactions on Systems, Man, and Cybernetics*, 2009.

[35] Thomas, L. C., "A Survey of Credit and Behavioral Scoring : Forecasting Financial Risk of Lending to Consumers", *International Journal of Forecasting*, Vol.16(2000), pp.149-172.

[36] Japkowicz, N. and S. Stephen, "The Class Imbalance Problem : A Systematic Vapnik, V.", *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[37] Wen, T. and A. Edelman, “A fast projected conjugate gradient algorithm for training support vector machines”, 2003.

[38] West, D., “Neural network credit scoring models”, *Computers and Operations Research*, 2000.

[39] Wolpert, D., “Stacked Generalization”, *Neural Networks*, Vol.5(1992), pp.241-259.

[40] Wu, G. and E. Y. Chang, “Class-Boundary Alignment for Imbalanced Dataset Learning”, *ICML*, 2003.

[41] Yang, C. Y., J. S. Yang and J. J. Wang, “Margin calibration in SVM class-imbalanced learning”, *Neurocomputing*, 2009.

[42] Yu, H., J. Yang and J. Han, “Classifying large data sets using SVMs with hierarchical clusters”, *Proceedings of the 9th ACM SIGKDD*, 2003.

◈ 저 자 소 개 ◈

김 진 우 (jinwkim@hongik.ac.kr)

현재 코리아크레딧뷰로(KCB, Korea Credit Bureau)에 근무하고 있으며, 홍익대 산업공학과 박사과정에서 데이터마이닝을 전공하고 있다. 국내 신용카드사들의 부정사용방지 시스템(FDS, Fraud detection System) 구축에 참여하였다. 관심 분야는 데이터마이닝, Risk Management 및 고객관계관리 등이며, 특히 관련 예측 모형의 개발 외에도 시스템 구축에 많은 관심을 가지고 있다.

지 원 철 (jhee@hongik.ac.kr)

현재 홍익대학교 산업공학과 교수로 재직 중이다. KAIST에서 지능정보시스템 전공으로 박사학위를 취득한 후, University of Illinois, Urbana-Champaign에서 초빙교수를 지냈다. 2000년대 초반 한국데이터마이닝학회 설립을 주도하였으며, 학회 회장을 역임하였다. Decision Support Systems, Expert Systems with Applications, Intelligent Systems in Accounting, Finance and Management 및 Neurocomputing 등의 국제학술지에 논문을 게재하였다. 관심 분야는 데이터마이닝, 빅 데이터 분석, 지능형 의사결정지원시스템으로 최근 신용카드 부정사용방지 시스템(Fraud Detection System)의 구축에 많은 관심을 가지고 있다.