지능정보연구 제18권 제2호 2012년 6월(pp.29~45)

회사채 신용등급 예측을 위한 SVM 앙상블학습

김명종
부산대학교 경영학과
(mjongkim@pusan.ac.kr)

회사채 신용등급은 투자자의 입장에서는 수익률 결정의 중요한 요소이며 기업의 입장에서는 자본비용 및 기업가치와 관련된 중요한 재무의사결정사항으로 정교한 신용등급 예측 모형의 개발은 재무 및 회계 분야에서 오랫동안 전통적인 연구 주제가 되어왔다. 그러나, 회사채 신용등급 예측 모형의 성과와 관련된 가장 중요한 문제는 등급별 데이터의 불균형 문제이다. 예측 문제에 있어서 데이터 불균형(Data imbalance) 은 사용되는 표본이 특정 범주에 편중되었을 때 나타난다. 데이터 불균형이 심화됨에 따라 범주 사이의 분류경계영역이 왜곡되므로 분류자의 학습성과가 저하되게 된다. 본 연구에서는 데이터 불균형 문제가 존재하는 다분류 문제를 효과적으로 해결하기 위한 다분류 기하평균 부스팅 기법 (Multiclass Geometric Mean-based Boosting MGM-Boost)을 제안하고자 한다. MGM-Boost 알고리즘은 부스팅 알고리즘에 기하평균 개념을 도입한 것으로 오분류된 표본에 대한 학습을 강화할 수 있으며 불균형 분포를 보이는 각 범주의 예측정확도를 동시에 고려한 학습이 가능하다는 장점이 있다. 회사채 신용등급 예측문제를 활용하여 MGM-Boost의 성과를 검증한 결과 SVM 및 AdaBoost 기법과 비교하여 통계적으로 유의적인 성과개선 효과를 보여주었으며 데이터 불균형 하에서도 벤치마킹 모형과 비교하여 견고한 학습성과를 나타냈다.

논문접수일 : 2011년 12월 01일 논문수정일 : 2012년 05월 29일 게재확정일 : 2012년 05월 30일
투고유형 : 국문일반 교신저자 : 김명종

1. 서론

기업이 영업활동을 수행하는 과정에서 필요한 장기 자금ㅊ의 조달 방법은 크게 주식 발행이나 부채 차입으로 구분할 수 있다. 주식 발행을 통한 자금조달은 경영의사결정에 의결권을 행사할 수 있기 때문에 경영권 방어에 부담이 될 수 있지만, 회사채(bonds) 발행을 통한 자금조달은 경영권에 대한 압박을 피할 수 있고 동시에 증권시장을 통하여 장기자금의 조달이 가능하므로 기업의 자금조달 방법으로 많이 활용된다. 또한 사채수요자(투자자)의 입장에서도 일정한 약정이율과 원금상환을 보증 받을 뿐만 아니라, 자본시장을 통하여 자금의 회수(현금화)가 용이하기 때문에, 회사채는 금융시장에서 자금의 배분기능을 수행한다고 볼 수 있다. 회사채는 1년 이상의 장기자금을 직접 금융시장에서 조달하는 채무증권이기 때문에, 발행기업에 대한 전문지식이 부족한 투자자를 보호할 체제가 우선되어야 한다. 이에 우리나라에서는 무보증사채에 한하여 발행 시 신용평가전문기관의 신용등급 첨부를 의무화하고 있다. 이런 맥락에서, 신용평가는 회사채가 거래되는 채권시장의 효

* 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 연구되었음(NRF-2010-332-B00090).

율성과 안정성을 도모하기 위하여, 개별투자자에게 중요한 투자등급 정보를 제공함으로써 발생 가능한 신용위험을 측정하고 통제할 수 있게 해주는 유용한 감독수단이며 회사채의 발행금리 및 주식가격과 연관되어 있기 때문에, 기업의 자본비용과 기업가치에 영향을 미친다고 할 수 있다. 이와 같이 기업의 신용등급은 투자자의 입장에서는 투하자본 수익률 결정의 중요한 요소이며 기업의 입장에서는 자본비용 및 기업가치와 관련된 중요한 재무의사결정사항으로 정교한 신용등급 예측모형의 개발은 재무 및 회계 분야에서 오랫동안 전통적인 연구 주제가 되어왔다.

회사채 신용등급 예측모형의 개발과 관련하여 기존 연구에서 많이 활용되어 왔던 기법은 크게 통계 모형과 인공지능 모형으로 구분할 수 있다. 회사채 등급평가 모형개발을 위하여 사용된 통계모형으로는 회귀분석, 다중판별분석(Multiple Discrimant analysis), 로지스틱 회귀분석, Probit 분석 등이 있다(Fisher, 1959; Pogue and Solofsky, 1969; West, 1991; Pinches and Mingo, 1973; Ederington, 1985; Gentry et al., 1988; Jackson and Boyd, 1988). 그러나, 통계적 모형은 선형성, 정규성 및 설명변수간 독립성 문제 등의 엄격한 통계적 가정으로 인하여 실제 활용에 상당한 제약을 받아왔다.

통계적 가정의 제약성을 해결하기 위하여 1980년대 후반부터 사례기반추론(Case-Based Reasoning : CBR), 의사결정트리(Decision Tree), 인공신경망(Neural Networks) 등 다양한 인공지능기법을 이용한 부실예측모형 및 회사채 등급평가 모형이 개발되어 왔다(옥중경 외, 2009; Dutta and Shekhar, 1988; Kwon et al., 1997; Maher and Sen, 1997; Chaveesuk et al., 1999; Huang et al., 2004). 이러한 실증연구들은 인공신경망의 예측성과가 가장 우수함을 보고하고 있다. 인공신경망과 더불어 최근 분류 및 예측 알고리즘의 성과개선과 관련하여 주목받고 있는 학습방법 중 하나는 Support vector Machine(SVM)이다. SVM은 Vapnik(1995)에 의해 제안된 분류 및 회귀학습 이론으로서 분류 문제에 있어 두 분류 사이의 거리인 마진(margin)을 최대로 하는 초평면(hyperplane)을 탐색하는 방법이다. 인공신경망과 비교하여 SVM은 첫째, 명료한 이론적 근거에 기반하므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어 높은 성과를 나타내고, 셋째, 입력변수의 차원에 의존하지 않고 자료의 수에 의존하여 신속하게 학습을 수행할 수 있으며, 넷째, 구조적 위험 최소화(Structural Risk Minimization)에 기반하므로 과대적합(Overfitting) 문제에 견고하다는 장점이 있다. 이러한 장점으로 인하여 SVM은 문자인식, 이미지 인식, 마이크로어레이 분석 등 자연과학 분야에서 활발하게 적용되어 왔다. 최근 SVM은 시계열 예측 및 분류(Cao and Tay, 2001; Kim, 2004; Tay and Cao, 2002), 기업부실예측(Shin et al., 2005; Min et al., 2006), 채권등급평가(안현철 외, 2006; 신택수 외, 2011) 등 경영분야에도 활발하게 도입되고 있으며 도입 결과, SVM의 예측성과가 다른 인공지능 기법이나 통계모형에 비하여 우수한 것으로 보고되고 있다.

회사채 신용등급 예측 모형을 개발 시에 예측 모형의 성과에 가장 중대한 영향을 미치는 요인은 데이터 불균형 문제이다. 카드사기적발(Fawcett and provost, 1997), 휴대폰 사기적발(Weiss, 2004), Response modeling(Shin and Cho, 2005), Remote Sensing (Bruzzone and Serpico, 1997), Scene Classification (Wu et al., 2003) 등의 분류 및 예측 문제에서 빈번히 관찰되는 데이터 불균형(Data Imbalance)은 사용되는 표본이 특정 범주에 편중되었을 때 나타난다. 회사채 신용등급 데이터 역시 특정 등급에 편중되는

데이터 불균형문제가 존재하며 이로 인하여 정교한 예측모형의 개발이 어렵게 된다. 불균형적인 데이터를 활용하여 예측모형을 구성하는 경우 파생되는 문제로서 1) 성과지표의 유효성 문제 및 2) 학습성과의 저하문제 등이 지적되고 있다(강필성 외, 2006; 김명종, 2009; Kotsiantis et al., 2007; Wang and Japkowicz, 2009).

SVM 역시 2개의 범주를 가진 이분류 문제를 해결하기 위한 이분류 기법으로 제안되었기 때문에 회사채 등급평가와 같은 다분류 문제(Mmulti-class Classification Problems)의 해결에는 많은 한계점을 가지고 있다. 이러한 한계점을 극복하기 위하여 One-Against-One 및 One-Against-All과 같은 다양한 학습기법들이 제안되어 왔으나, 현재까지 이분류 문제에서 보여준 탁월한 성과를 보여주지 못하고 있다. 안현철 등의 연구(2006)에서는 이러한 문제를 해결하기 위하여 다분류 SVM을 활용한 기업채권평가모형을 개발하였으며, 국내채권등급 자료를 이용하여 다분류 SVM의 성과를 검증한 결과, 다중판별분석과는 1% 수준에서, 인공신경망과는 10% 수준에서 다분류 SVM의 성과가 우수함을 보여주었다. 그러나 이들의 연구에서는 단순평균정확도 개념에 기초한 성과차이만을 보여주었을 뿐 등급별 데이터 불균형을 고려한 정확도 개념을 고려하지 못하였다.

기계학습 분야에서 데이터 불균형문제의 해결에 활용되어 왔던 대표적인 기법은 부스팅 알고리즘이다. 부스팅 알고리즘은 순차적 분류자 생성기법으로 이전 분류자에서 오분류된 관측치에 높은 가중치를 부여하고 새로운 분류자 생성을 위한 학습표본의 추출 시점에서 가중치가 높게 부여된 오분류 관측치들이 새로운 학습표본에 많이 포함되기 때문에 부스팅 알고리즘은 오분류 관측치에 초점을 맞춘 학습을 진행할 수 있게 되며, 따라서 오분류율이 높은 소수 범주 표본에 대한 학습을 강화할 수 있다는 장점이 있다. 많은 이론적 연구를 통하여 부스팅 알고리즘 기반의 SMOTEBoost(Chawla et al., 2003), RUSBoost(Seiffert et al., 2008), AdaBoost(Freund and Schapire, 1997), Geometric Mean-based Boosting(GM-Boost)(김명종, 2009) 등이 제안되어 왔으며, 실증 연구를 통하여 부스팅 알고리즘 기반의 기법들이 통계적 샘플링 기법보다 데이터 불균형 문제의 해결에 효과적임을 보여주었다(김명종, 2009; Wang and Japkowicz, 2009). 이 중 GM-Boost 알고리즘은 전통적으로 활용되어 왔던 AdaBoost 알고리즘에 기하평균 개념을 도입한 새로운 부스팅 알고리즘으로 다수 범주(정상기업)과 소수 범주(부실기업) 사이에 불균형 분포를 보이는 기업부실 예측문제에 GM-Boost 기법을 적용하였다. 분석 결과, GM-Boost 기법은 통계적 기법 및 기존의 부스팅 알고리즘 기법과 비교하여 유의적인 성과차이가 나타나며, 범주간의 데이터 불균형 비율에 관계없이 높은 정확성과 견고한 학습능력을 보유하고 있음을 보여주었다(김명종, 2009). 그러나, GM-Boost 알고리즘은 이범주분류에만 활용이 가능하다는 단점이 있다.

본 연구에서는 이범주 분류문제에만 활용 가능하였던 기존의 GM-Boost를 확장하여 다범주 분류문제를 해결하기 위한 다분류 GM-Boost(Multiclass GM-Boost : MGM-Boost) 알고리즘을 제안하고자 한다. 기존의 GM-Boost는 2개 범주의 분류문제에 한정되어 적용된 반면, 본 연구에서 제안하고자 하는 MGM-Boost는 이분류 문제뿐만 아니라 다분류 문제 해결에 적용이 가능한 일반화된 데이터 불균형 해결기법이라 할 수 있다. 본 연구에서는 대표적인 다분류 문제인 회사채 등급평가 문제에 적용하여 제안 알고리즘의 성과를 검증하고자 한다. 실증분석 결과, MGM-Boost 알고리즘은 SVM 및 AdaBoost 알고리즘과 비교하여 단순평균 정확도뿐만 아니라

가중평균 정확도 측면에서도 우수한 정확성을 보유하고 있으며, 통계적으로 유의적인 성과차이를 보여주었다.

본 연구는 다음과 같이 구성되어 있다. 제 2장에서는 회사채 등급평가 문제와 관련된 데이터 불균형 문제를 검토하고자 한다. 제 3장에서는 본 연구에서 제안한 MGM-Boost 알고리즘에 대하여 설명하고자 한다. 제 4장에서는 제안 모형의 유용성을 확인하기 위한 실험 데이터 수집 및 실험 설계 과정에 대하여 설명한다. 제 5장에서는 GM-Boost의 성과 검증 결과를 제시하고자 한다. 마지막 제 6장에서는 결론과 함께 미래연구방향을 제시하고자 한다.

2. 회사채 등급평가의 데이터 불균형 문제

<Table 1>에서는 2008년부터 2011년 9월 말까지의 국내 신용평가전문기관의 회사채 등급자료를 보여주고 있다. 20011년 9월 말 기준으로 AAA등급(16.0%), AA등급(32.3%) 및 A등급(31.7%)에 대하여 신용등급의 약 80%가 3개 등급에 편중되어 있다. 특히 BBB등급 이상의 투자등급 편중도는 약 90.3%이고 BB등급 이하의 투기등급 편중도는 약 9.7%로 회사채 신용등급 자료에 심각한 데이터 불균형이 존재하고 있음을 보여주고 있다.

이러한 불균형적인 데이터를 활용하여 예측모형을 구성하는 경우 파생되는 문제로서 1) 성과지표의 유효성 문제 및 2) 학습성과의 저하문제가 지적되고 있다.

2.1 성과지표의 유효성

현재까지 분류자의 성과 측정에 보편적으로 활용되는 지표는 단순평균 정확도로 전체 표본 중 정분류된 표본의 비율로 계산된다. 예를 들어 m개의 범주를 가진 n개의 관측치를 가정하면 단순평균 정확도는 다음과 같이 계산된다. 여기에서 $P_j(i)$는 j번째 실제범주에 속한 i번째 관측치를 j번째 범주로 정확히 예측하는 경우 1이 되며 부정확하게 분류하는 경우 0이된다.

$$\sum_{j=1}^{m} p_j(i) \Big/ _n \quad (i=1,\ 2,\ \cdots,\ n,\ \ j=1,\ 2, \cdots,\ m)$$

$$\text{where} \quad p_j(i) = \begin{cases} 0 & p_j(i) \neq y_i \\ 1 & p_j(i) = y_i \end{cases}$$

<Table 1> The Sample Observations Across Ratings

Rating	2008. 12. 31	2009. 12. 31	2010. 12. 31	2011. 9. 31
AAA	44(14.1%)	50(15.0%)	51(15.0%)	56(16.0%)
AA	60(19.2%)	87(26.1%)	102(30.0%)	113(32.3%)
A	86(27.6%)	91(27.3%)	109(32.1%)	111(31.7%)
BBB	50(16.0%)	42(12.6%)	36(10.6%)	36(10.3%)
BB	24(7.8%)	19(5.7%)	17(5.0%)	17(3.1%)
B	35(11.2%)	26(7.8%)	16(4.7%)	16(4.6%)
Below B	13(4.2%)	18(5.4%)	9(2.6%)	9(2.0%)
Investment Grade	240(76.9%)	270(81.1%)	298(87.6%)	316(90.3%)
Speculation Grade	72(23.1%)	63(18.9%)	42(12.4%)	34(9.7%)
Total	312	333	340	350

Source : Bond rating data of Korea Information Service Co.

그러나 데이터 불균형이 존재하는 상황에서 단순평균 정확도는 다수 범주(Majority Class) 표본의 분류 정확성에 의존하여 분류자의 성과를 왜곡하는 단점이 있다. 예를 들어 <Table 1>과 같은 회사채 등급평가 자료를 활용하는 경우, 단순평균 정확도는 다수 범주인 AA 및 A등급의 분류 정확성에 의존하여 지속적으로 증가하지만, 정작 소수 등급인 투기등급의 분류 정확성은 감소하게 되므로 분류자의 성과측정에 상당한 왜곡현상이 발생된다. 결과적으로 단순평균 정확도는 균형적인 분포를 보이는 데이터에 대한 분류자의 성과 지표로 적합할 수 있지만, 데이터 불균형 하에서는 다수 범주의 분류 정확성에 의존하여 분류자의 성과를 결정하기 때문에 더 이상 적합한 성과 지표가 되지 못한다(강필성 외, 2006; 김명종, 2009; Kotsiantis et al., 2007; Wang and Japkowicz, 2009).

이러한 문제를 해결하기 위하여 기하평균 정확도(Geometric-Mean Accuracy)가 단순평균 정확도를 대체하여 이용되고 있다. 범주별 정확도를 동시에 고려한 성과지표인 기하평균 정확도는 $\sqrt[m]{\prod_{j=1}^{m} p_j(i)}$ $(i=1, 2, \cdots, n,\ j=1, 2, \cdots, m)$로 계산된다(Kubat et al., 1997).

2.2 학습성과 저하 문제

데이터 불균형으로 파생되는 두 번째 문제는 분류자의 학습 성과가 저하되는 문제이다. 데이터 불균형 하에서는 다수 범주 표본에 의한 분류 경계영역의 침해로 인하여 소수 범주 영역이 점차로 축소하고 결과적으로 소수 범주에 대한 분류 정확성이 급격히 감소된다.

강필성 등의 연구(2006)에서는 이분류 문제를 대상으로 데이터 불균형이 SVM의 분류 정확성에 미치는 영향을 분석하였다. 이들은 데이터 균형 비율에 따라 6개의 표본 집합(1 : 1, 1 : 3, 1 : 5, 1 : 10, 1 : 30, 1 : 50)을 구성하고 SVM을 이용한 분류 실험을 수행하였다. 연구 결과 불균형 비율이 크지 않은 표본 집합(1 : 1, 1 : 3)의 경우 두 범주 사이의 경계 영역의 크기가 유사함을 보여주었다. 그러나 불균형 비율이 심해진 표본집합(1 : 5, 1 : 10)의 경우 다수 범주의 표본이 소수 범주의 영역을 침범하게 되어 소수 범주의 영역이 점점 작아지기 때문에 소수 범주에 속하는 표본의 분류 정확성이 감소하는 것을 확인하였다. 특히, 극단적인 불균형을 보이는 표본집합(1 : 30, 1 : 50)의 경우 분류자의 소수 범주에 대한 영역이 과도하게 작아져 소수 범주에 대한 분류 자체가 큰 의미가 없음을 보고하고 있다. 또한 데이터 불균형이 심해질수록 소수 범주 표본의 분류 정확도가 크게 감소하고 이에 따라서 기하평균 정확도는 감소하지만 단순평균 정확도는 오히려 다수 범주의 높은 분류 정확도에 의존하여 꾸준히 증가함을 보여주었다. 이러한 결과를 기초로 데이터 불균형 상황에서 단순평균 정확도는 성과 지표로서 적합하지 않음을 주장하였다.

Wu and Chang(2003)은 데이터 불균형으로 인한 SVM의 경계영역의 왜곡(Skewed Boundary)의 원인을 다음 두 가지로 보고하고 있다. 첫째, 학습데이터 비율의 불균형으로 소수 범주 표본이 소수 범주 경계영역 내에 존재하지 않으려는 경향이 발생한다. 둘째, Support Vector 비율의 불균형으로 다수 범주에 과도한 표본이 집중되는 경우 다수 범주의 분류 경계영역이 확대되고 소수 범주의 분류 경계영역이 축소되는 경계영역의 왜곡이 나타나며, 결과적으로 분류자는 예측 표본을 다수 범주로 분류할 가능성이 높아지게 된다는 분석 결과를 제시하였다.

2.3 SVM의 다분류 문제 해결기법

SVM은 이분류 문제 해결을 위하여 제안된 기법으로 회사채 등급평가와 같은 다분류 문제에 대한 한계점이 지적되어왔다. 이러한 문제의 해결 대안으로서 회사채 등급평가 연구에서 보편적으로 이용되고 있는 방법은 1) 이분류 SVM을 여러 개 결합하는 방법과 2) 모든 Class를 한번에 고려하여 하나의 최적화 문제로 해결하는 방법으로 구분 할 수 있다.

첫 번째 방법으로 (1) One-against-all 방법과 (2) One-against-one 방법이 제안되었다. 그러나, 위의 두 가지 방법에 기반한 SVM 학습은 이분류 문제와 같은 우수한 예측 정확성을 다분류 문제에서는 보여주지 못하고 있다. 한편 두 번째 방법의 경우 최종 수정된 분리 경계면을 해당 문제의 Class에 맞추어 분류할 수 있어야 하므로 이차계획문제에 포함된 선형제약식이 해당 문제의 Class 만큼 늘어나게 된다. 이로 인하여 이차계획문제의 해를 찾는데 있어서 이러한 다분류 SVM 모형은 일반적인 이분류 SVM 모형에 비하여 보다 훨씬 복잡한 분해방식을 요구하게 된다. 이러한 문제에 대한 대안으로 이차계획문제의 목적식에 특정한 항을 임의로 추가해 분해를 용이하게 하는 방법과 여유변수(slack variable)를 활용한 방법이 제안되었다(Hsu and Lin, 2002). 이러한 방법을 이용하여 회사채 등급평가 모형개발에 활용된 연구들은 SVM이 기타 분류기법보다 예측성과가 우수함을 보고하고 있다(안현철 등, 2006) 그러나, 이러한 방법을 수행하기 위하여 SVM 모형들은 대량의 수학적 계산을 수행해야 하며, 학습시간을 줄이기 위하여 다양한 Approximation algorithms(decomposition methods, sequential minimal optimization algorithm)을 사용하고 있지만, 이로 인하여 분류 성과가 저하될 수 있음이 지적되고 있다.

3. MGM-Boost 알고리즘

부스팅 알고리즘은 소수 범주 표본에 대한 학습을 효과적으로 진행할 수 있다는 장점으로 인하여 SMOTEBoost(Chawla et al., 2003), RUSBoost(Seiffert et al., 2008), AdaBoost(Freund and Schapire, 1997), GM-Boost(김명종, 2009) 등 다양한 부스팅 알고리즘(Boosting algorithms)이 데이터 불균형 문제의 해결 대안으로 제안되고 있다. 본 장에서는 대표적 부스팅 기법인 AdaBoost 알고리즘과 본 연구에서 제안한 MGM-Boost 알고리즘에 대하여 비교하여 설명하고자 한다.

3.1 AdaBoost 알고리즘

앙상블 학습은 최근 기계학습 분야에서 분류자(classifier)의 정확도 개선을 위하여 제안된 기계학습기법이다. 앙상블 학습은 학습 데이터를 정확하게 표현할 수 있는 하나의 가설을 선택하는 것이 아니라 가설들의 집합을 구성하고 새로운 데이터를 예측할 때에는 그 가설들의 예측을 결합하여 최종 결정을 내린다. 이를 위하여 앙상블 학습은 복수의 기저 분류자(Base Classifiers) 집합인 앙상블을 구성하고, 앙상블에서 추론된 복수의 학습결과를 단일 강분류자(A Single Strong Classifier)를 통하여 결합한 최종 결과를 산출하게 된다.

AdaBoost 알고리즘은 앙상블 학습 알고리즘 중 가장 일반적으로 사용되고 있는 부스팅 알고리즘으로 Freund and Schapire(1997)에 의하여 제안되었다. 부스팅은 임의 추측보다 다소 높은 수준의 정확성을 보유한 여러 개의 약분류자의 선형결합으로 정확성이 높은 강분류자를 생성하는 알고리즘이다. 부스팅 방법은 이전 분류자의 성과를 기초로 오분류된 관측치에 초점을 맞추어 분류자를 순차적

으로 생성한다. AdaBoost 알고리즘의 설명을 위하여 m개의 범주를 가진 n개의 학습 표본과 B개의 기저 분류자로 구성된 앙상블 $C = \{C_1, C_2, \cdots, C_B\}$을 가정하면 b번째 기저 분류자의 오류율(e_b)은 다음과 같이 계산된다.

$$\varepsilon_b = \sum_{i=1}^{n} \omega_b(i)\xi_b(i)$$
$$\text{where} \quad \xi_b(i) = \begin{cases} 1 & C_b(x_i) \neq y_i \\ 0 & C_b(x_i) = y_i \end{cases}$$

여기에서 $w_b(i)$는 i번째 관측치에 부여되는 가중치로 초기 가중치는 $1/n$로 설정된다. 또한 xi는 i번째 관측치의 예측변수 벡터이고 yi는 i번째 관측치의 실제범주를 나타내며 $C_b(x_i)$는 예측변수 벡터 xi에 대한 b번째 분류자의 예측결과이다. b+1번째 분류자에서 i번째 관측치에 부여되는가중치는 $w_{b+1}(i) = w_b(i)\exp(\alpha_k \xi_b(i))$로 조정되어 오분류된 관측치에 더 높은 가중치가 부여된다. 여기에서 b는 분류자의 중요도 또는 정확도의 개념으로 해석되며 $\alpha_k = 1/2\ln(1-\varepsilon_b/\varepsilon_b)$로 계산된다. b+1번째 분류자의 학습표본을 구성할 때 가중치가 높은 오분류 관측치가 많이 포함되기 때문에 부스팅 알고리즘은 오분류 관측치에 초점을 맞춘 학습을 진행할 수 있게 된다. 이러한 방식으로 새로운 분류자 b = 1, 2, 3, B가 생성되며 i번째 관측치의 최종결과는 다음과 같이 각 앙상블의 결과의 가중평균으로 계산된다.

$$C(x_i) = \arg_{yi} Max \sum_{b=1}^{B} a_b C_b \delta(C_b(x_i),\ y_i)$$
$$= \arg_{yi} Max \sum_{b:\, C_b(x_i) = y_i}^{B} a_b$$

소수 범주 표본에 대한 학습기회를 제공한다는 장점으로 인하여 AdaBoost 알고리즘을 기반으로 한 다양한 부스팅 알고리즘이 데이터 불균형 문제의 해결 대안으로 자주 활용되고 있다. 데이터 불균형이 심할수록 소수 범주에 대한 오류율은 높게 나타나는 반면, 다수 범주에 대한 오류율은 낮게 나타나게 된다. 결과적으로 새로운 분류자의 학습표본을 추출하는 과정에서 높은 가중치가 부여된 소수 범주 표본들이 새로운 학습 표본에 많이 포함되므로 새로운 분류자는 소수 범주에 대한 학습을 강화하게 된다. 이러한 방식으로 학습 초기에는 다수 범주에 편중된 표본 학습에서 시작되더라도 순차적으로 소수 범주 표본에 대한 학습기회가 많아지게 된다. 이러한 특성으로 인하여 부스팅 알고리즘은 데이터 불균형 하에서도 견고한 학습성과를 나타낼 수 있다는 장점이 있다.

그러나 부스팅 알고리즘은 단순 평균 개념에 기초한 주요 파라미터로 인하여 다음과 같은 문제가 나타날 수 있다. 첫째, 분류자의 오류율 e_k는 단순평균 오류율로 전체 표본 대비 오분류 표본 비율로 계산된다. 불균형 데이터의 경우 다수 범주의 낮은 오류율로 인하여 단순평균 오류율이 왜곡될 수 있다. 둘째, 분류자의 성과를 나타내는 b 역시 단순평균 정확도에 기초한 개념이다. 이미 언급한 바와 같이 데이터 불균형 하에서 단순평균 정확도는 성과지표로 유효하지 않기 때문에 범주별 데이터 불균형을 고려한 가중평균 정확도 개념으로 대체할 필요가 있다.

3.2 MGM-Boost 알고리즘

앞서 언급한 AdaBoost 알고리즘의 기본 가정과 더불어 m개의 범주를 가진 n개의 표본이 b번째 분류자의 학습표본으로 추출되었다고 가정하면 b번째 분류자의 j번째 범주의 오류율을 ε_b^j는 다음과 같이 계산된다. 여기에서 $\omega_b^j(i)$는 j번째 범주에 속하

는 i번째 관측치에 부여되는 가중치로 AdaBoost와 마찬가지로 초기 가중치는 $1/n$로 설정된다. $C_b^j(x_i)$는 j번째 범주에 속하는 예측변수 벡터 xi에 대한 b번째 분류자의 예측결과이다.

$$\epsilon_b^j = \sum_{i=1}^{n} \omega_b^j(i)\,\xi_b^j(i)$$

$$\text{where} \quad \xi_b^j(i) = \begin{cases} 1 & C_b^j(x_i) \neq y_i \\ 0 & C_b^j(x_i) = y_i \end{cases}$$

범주별 오류율을 기초로 b번째 기저 분류자의 오류율(eb)은 다음과 같이 기하평균 오류율 $\varepsilon_b = \sqrt[m]{\prod_{j=1}^{m} \varepsilon_b^j}$로 계산된다. b+1번째 분류자에서 j번째 범주에 속하는 i번째 관측치에 부여되는 가중치는 $\omega_{b+1}^j(i) = \omega_b^j(i)\exp(\alpha_b^j \xi_b^j(i))$로 조정되어 오분류율이 높은 범주와 더불어 오분류된 관측치에 더 높은 가중치가 부여된다. 여기에서 α_b^j는 $\alpha_b^j = 1/2\ln(1-\varepsilon_b^j/\varepsilon_b^j)$로 계산된다. 이러한 방식으로 새로운 분류자 b = 1, 2, 3, …, B가 생성되며 분류자의 분류 정확성을 의미하는 b는 범주별 정확도의 가중평균 정확도로 $\alpha_b = \sqrt[m]{\prod_{j=1}^{m} \alpha_b^j}$로 계산된다. i번째 관측치의 최종결과는 다음과 같이 각 앙상블의 결과의 가중평균으로 산출된다.

$$C(x_i) = \arg_{yi} Max \sum_{b=1}^{B} a_b C_b \delta(C_b(x_i), y_i)$$
$$= \arg_{yi} Max \sum_{\substack{b: C_b(x_i)=y_i \\ i}}^{B} \alpha_b$$

MGM-Boost 알고리즘은 AdaBoost 알고리즘에 기초하고 있으므로 소수 범주 표본에 대한 학습기회를 강화할 수 있다는 장점을 가지고 있으며 가중평균 오류율과 가중평균 정확도 개념을 사용하고 있으므로 범주별 분류 정확도를 동시에 고려한 학습을 진행할 수 있다는 장점이 있다. <Figure 1>은 GM-Boost 알고리즘의 절차를 간략하게 요약하고 있다.

4. 연구 설계

4.1 데이터 수집 및 변수선정

본 연구에서는 회사채 등급평가 자료를 이용하

1. Start with $w_1^j(i) = 1/n,\ (i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m)$
2. Repeat for creating $C_b,\ b = 1, 2, \cdots, B)$
   a) Fit the classifier $C_b(x_i) \in \{1, 2, \cdots, m\}$
   b) Compute geometric error rate
   $\epsilon_b = \sqrt[m]{\prod_{j=1}^{m} \varepsilon_b^j}$
   where $\varepsilon_b^j = \sum_{i=1}^{n} \omega_b^j(i)\,\xi_b^j(i),\ \xi_b^j(i) = \begin{cases} 1 & C_b^j(x_i) \neq y_i \\ 0 & C_b^j(x_i) = y_i \end{cases}$
   c) Compute geometric average accuracy
   $\alpha_b = \sqrt[m]{\prod_{j=1}^{m} \alpha_b^j}$ where $\alpha_b^j = 1/2\ \ln(1-\varepsilon_b^j/\varepsilon_b^j)$
   d) Weights adjustment against data of each class $\omega_{b+1}^j(i) = \omega_b^j(i)\exp(\alpha_b^j \xi_b^j(i))$
3. Output the final classifier
   $C(x_i) = \arg_{yi} Max \sum_{b=1}^{B} a_b C_b \delta(C_b(x_i), y_i) = \arg_{yi} Max \sum_{\substack{b: C_b(x_i)=y \\ i}}^{B} \alpha_b$

<Figure 1> MGM-Boost Algorithm

여 본 연구에서 제안한 MGM-Boost 알고리즘의 성과를 검증하고자 한다. 성과검증을 위하여 2003~2006년의 2100개 제조기업의 회사채 등급자료를 국내 신용평가기관에서 수집하였다. 수집된 회사채 등급자료의 등급별 분포는 <Table 2>에 제시되어 있다. 일반적으로 회사채의 신용등급은 신용도에 따라 크게 10개 등급(AAA, AA, A, BBB, BB, B, CCC, CC, C, D)으로 운영되고 있는데 본 연구에서는 5개 등급으로 구분하여 1(AAA, AA), 2(A), 3(BBB), 4(BB), 5(CCC, CC, C, D)로 표기하였다. 5개 등급으로 구분한 이유는 AAA 및 B등급 이하에 해당하는 회사채 자료가 거의 없어 실제적인 분석이 불가능하기 때문이다.

<Table 2> The Transformation Ratings Letters and Sample Observation Across Raings

Variables	Ratings	Obs.	Percentage
1	AAA, AA	231	11.00%
2	A	396	18.86%
3	BBB	728	34.67%
4	BB	564	26.86%
5	B, CCC, CC, C, D	181	8.62%
Total		2100	100%

회사채 등급예측을 위한 재무자료는 해당 기업의 등급 산출에 활용된 2001~2005년의 결산 재무자료를 활용하였다. 회사채 등급평가에 사용되는 재무비율은 일차적으로 기존의 회사채 등급평가 연구에 사용된 비율 및 실무에서 등급평가 예측지표로 유용하게활용되는 비율을 중심으로 30개의 재무비율을 수집하였으며 요인분석을 통하여 수익성, 부채상환능력, 레버지리, 자본구조, 유동성, 활동성 및 규모의 7개 재무비율 군으로 재분류하였다. 요인별 재무비율은 <Table 3>에 제시되어 있다.

최종 입력변수는 일요인 분산분석을 수행하여 각 요인별로 가장 큰 F 값을 가지고 있는 7개 재무비율을 선정하였다. 비록 변별력과직접적인 관련성은 없으나, 다중공선성 문제는 모형 개발 시 필수적으로 고려해야 할 문제이다. 본 연구에서는 7개 재무비율 사이의 다중공선성의 존재여부를 확인하기 위하여 분산팽창요인(Variance Inflation Factors : VIF) 분석을 실시하였다. 다중공선성이 존재한다고 의심되는 VIF 임계치는 5~10사이이며 VIF가 10 이상이면 다중공선성이 심각한 것으로 판단할 수 있다. <Table 4>에는 최종 선정된 7

<Table 3> Financial Ratios

Category	Variable	Category	Variable
Profitability	Ordinary income to total assets Net income to total assets (ROA) Financial expenses to sales Financial expenses to total debt Net financing cost to sales Ordinary income to sales Net income to sales Ordinary income to capital Net income to capital	Leverage	Capital to total asset Current assets to total assets
		Capital structure	Retained earning to total assets Retained earning to total debt Retained earning to current assets
		Liquidity	Cash ratio Quick ratio Current assets/current Liabilities
Debt coverage	EBITDA to Interest expenses EBIT to Interest expenses Cash operating income to interest expenses Cash operating income to total debt Cash flow after interest payment to total debt Cash flow after interest payment to total debt Debt repayment coefficient Borrowings to Interest expenses	Activity	Inventory to sales Current liabilities to sales Account receivable to sales
		Size	Total assets Sales Fixed assets

<Table 4> The results of F-test and VIF

Financial Ratio	F-Value	VIF
Ordinary income to total assets	25.683*	1.36
EBITDA to Interest expenses	17.888*	2.11
Capital to total asset	16.424*	1.77
Retained earning to total assets	23.882*	2.53
Cash ratio	9.139*	1.34
Inventory to sales	7.738*	1.59
Total assets	5.655*	1.31

Source : Represent Significance Levels at 1%.

개재무비율의 F-값 및 VIF 값이 제시되어 있으며, VIF 분석결과는 7개 변수의 VIF는 1.31~2.53으로 7개 변수 사이에 다중공선성이 실질적으로 존재하지 않음을 나타내고 있다.

4.2 실험설계

본 연구의 회사채 등급 예측모형 개발에 활용되는 기법은 1) SVM, (2) AdaBoost 알고리즘, (3) MGM-Boost 알고리즘이다. 기저 분류자(base classifier)로 활용되는 SVM은 Platt(1998)에 의해 제안된 다분류 SMO(Sequential Minimal Optimization) 알고리즘을 사용하였다. 본 연구에서는 SVM의 커널함수로서 가장 일반적으로 이용되는 가우시안 RBF를 사용하였다. 가우시안 RBF를 커널함수를 사용하기 위해서는 커널함수의 상한 C와 커널 파라미터 d^2를 고려해야한다. Tay and Cao(2001)에 따르면 적절한 C의 값으로 적절한 범위는 10에서 100사이이고 d^2의 적절한 범위는 1~100사이임을 보여주었다. 이러한 연구 결과에 기초하여 각각 10~100과 1~100의 범위 내에서 다양한 값을 대입하여 다양한 모형을 생성시켰다.

AdaBoost 알고리즘과 MGM-Boost 알고리즘은 분류자 생성 횟수가 25회를 넘어서면 오류 감소효과가 미미하다는 연구 결과(Opitz and Maclin, 1999)에 기초하여 최대 앙상블 생성횟수를 25회로 제한하였다.

4.3 연구결과

모형 적합과 모형의 성능평가에 모두 같은 자료를 써서 얻어지는 오분류율을 겉보기 오분류율(Apparent Misclassification Rate)라고 하며 겉보기 오분류율을 이용하는 경우 참오분류율(True Misclassification Rate)을 과소 추정한다고 알려져 있다. 이를 해결하기위하여 가장 널리 쓰이는 분석방법은 교차타당성에 의한 추정방법으로 보통 주어진 자료를 10등분하는 10-fold 검증방법(10-fold cross validation)을 일반적으로 사용하고 있다(Optiz and Macline, 1999).

<Table 5> Comparison of Predictive Accuracy

Variable	Obs.	Percentage	SVM	AdaBoost	MGM-Boost
1	231	10.99%	16.50%	18.20%	21.21%
2	396	18.84%	9.10%	19.11%	21.72%
3	728	34.73%	82.80%	82.10%	82.55%
4	564	26.83%	63.80%	63.80%	64.36%
5	181	8.61%	1.10%	5.00%	7.18%
Total	2100	100%			
AAA*			49.47%	51.69%	52.95%
AGA*			15.42%	24.65%	28.12%

Note : * AAA and AGA stands for Average Arithmetic Accuracy and Average Geometric Accuracy, respectively.

본 연구에서는 각 모형간 성과차이가 우연한 결과가 아님을 확인하기 위하여 10-fold 검증을 3회 반복 수행하였다. 이를 위하여 전체 2,100개의 기업을 표본수가 동일한 10개 fold로 구성하고 9개의 집합은 분석용 데이터로 활용하고 나머지 1개 fold를 검증용 데이터로 활용하게 된다. 이러한 방법으로 30회의 교차타당성 검증을 수행하였다. 30회 교차타당성의 검증결과는 <Table 5>에 제시되어 있다.

30회 교차타당성 검증 결과 각 모형의 단순평균 정확도는 MGM-Boost(43.81%~62.86%), AdaBoost(43.33%~62.16%), SVM(39.52%~60.48%)로 나타났다. 30개 검증표본에 대한 각 모형의 단순평균 정확도의 평균은 MGM-Boost(52.95%), AdaBoost(51.69%), SVM(49.47%)으로 GM-Boost 알고리즘은 다른 알고리즘에 비하여 1.26%~3.49% 정도 우수한 것으로 나타나고 있다. 각 모형의 기하평균 정확도는 MGM-Boost(19.52%~46.88%), AdaBoost(17.51%~41.05%), SVM(13.27%~34.94%)이며 평균은 MGM-Boost(28.12%), AdaBoost(24.65%), SVM(15.42%)로 나타났다. MGM-Boost 알고리즘과 다른 모형에 비하여 약 3.47%~12.7% 우수한 것으로 파악되었으며 단순평균 정확도의 성과차이에 비하여 기하평균의 성과차이가 더욱 커진 것으로 분석되었다.

등급별 정확도 관점에서 다수 범주에 속하는 3 및 4등급의 예측결과는 각각SVM(82.80%, 63.80%), AdaBoost(82,10%, 63,80%), MGM-Boost(82.55%, 64.36%)로 상당히 높은 수준의 예측정확성을 보여주었으나, 각 모형 사이의 뚜렷한 성과차이는 나타나지 않았다. 그러나, 소수범주인 1, 2 및 5등급에 대해서는 SVM(16.50%, 9.10%, 1.10%), AdaBoost(18.20%, 19.11%, 5.00%), MGM-Boost(21.21%, 21.72%, 7.18%)로 AdaBoost의 경우 SVM에 비하여 소수 범주의 예측력이 약 1.10~4.55배 개선되었고 MGM-Boost의 경우 SVM에 비하여 소수 범주의 예측력이 약 1.29~6.53배 개선된 것으로 나타났다. 결과적으로 MGM-Boost는 데이터 불균형하에서 다수범주에 대하여 SVM 및 AdaBoost와 유사한 수준의 예측력을 유지하는 반면 소수범주에 대하여 상당한 성과개선의 효과가 있는 것으로 분석되었다.

<Table 6> The Results of T-test

Classifier	AAA**		AGA**	
	AdaBoost	MGM-Boost	AdaBoost	MGM-Boost
SVM AdaBoost	3.114*	3.998* 2.882*	3.753*	4.995* 2.995*

Note : * represent significance levels at 1%.
** AAA and AGA stands for Average Arithmetic Accuracy and Average Geometric Accuracy, respectively.

이러한 예측력 차이의 유의성을 검증하기 위하여 T-test를 실시하였으며 그 결과는 <Table 6>에 제시되어 있다. T-test 결과에서는 AdaBoost 알고리즘은 SVM과 비교하여 1% 수준에서 유의적인 성과차이를 보여주었으며, MGM-Boost 역시 SVM과 AdaBoost 알고리즘과 비교하여 1% 수준에서 유의적인 성과차이를 나타냈다.

5. 결론 및 향후 연구 방향

데이터 불균형 문제는 분류자의 성과에 미치는 영향이 크기 때문에 패턴 인식과 기계학습 분야에서 관심을 받고 있는 이슈 중 하나이다. 본 연구는 데이터 불균형이 심화되는 환경에서도 높은 성과를 창출할 수 있고 견고한 분류자를 생성할 수 있는 MGM-Boost 알고리즘을 제안하였다. 회사채 신용등급 평가문제를 대상으로 MGM-Boost 알고리즘의 성과를 확인한 결과 MGM-Boost 알고리즘은 데이터 불균형하에서 다수범주에 대해서는 SVM과 유사한 수준의 예측력을 유지하는 반면 소수범주에 대해서는 상당한 성과개선의 효과가

있는 것으로 분석되었다. 결과적으로 MGM-Boost 알고리즘은 데이터 불규형하에서 다른 알고리즘에 비하여 높은 분류 정확성과 견고한 학습능력을 확보하고 있음을 확인하였다.

본 연구와 관련하여 다음과 같은 미래연구가 지속되기를 기대한다.

첫째, 부스팅 알고리즘은 학습표본에 이상치(Outlier)를 가진 특정 관측치가 포함되거나 앙상블 분류자 사이의 상관관계가 높은 경우 분류 정확도가 감소되는 문제가 발생하는 단점이 있다(Optiz and Maclin, 1999). 이러한 단점을 보완하기 위하여 다양한 방법(Maia et al., 2009; Cover and Thomsa, 1991; Darbellay, 1999) 들이 제안되고 있으며 후속 연구에서는 이러한 방법과 결합된 알고리즘을 개발 연구를 수행하고자 한다.

둘째, 본 연구에서 제안한 앙상블 기법은 부스팅 알고리즘의 수정을 통하여 데이터 불균형 문제를 해결하는 방향으로 진행되었다. 그러나 본 연구의 결과를 SVM의 커널조정과 연계하는 방법으로 데이터 불균형 문제를 해결할 수 있기 때문에 이러한 후속연구가 진행되길 기대한다(Hong, 2007; Wu et al., 2005).

참고문헌

강필성, 조성준, “데이터 불균형 해결을 위한 Under-sampling 기반 앙상블 SVMs”, 대한산업공학회/한국경영과학회 2006 춘계공동학술대회, 2006.

김명종, “기업부실 예측 데이터의 불균형 문제 해결을 위한 앙상블 학습”, *지능정보연구*, 2009, 15권 3호(2009), 1~15.

신택수, 홍태호, “AdaBoost 알고리즘 기반 SVM을 이용한 부실 확률분포 기반의 기업신용평가”, *지능정보연구*, 17권 3호(2011), 25~41.

안현철, 김경재, 한인구, “다분류 Support Vector Machine을 이용한 한국기업의 지능형 기업채권 평가모형”, *경영학연구*, 35권 5호(2006), 1479~1496.

옥중경, 김경재, “유전자 알고리즘 기반의 기업부실예측 통합모형”, *지능정보연구*, 15권 4호(2009), 99~121.

Bruzzone, L. and S. B. Serpico, “Classifications of imbalanced remote-sensing data by neural networks”, *Pattern recognition letters*, Vol.18, No.11~13(1997), 1323~1328.

Cao, L. and F. E. H. Tay, “Financial forecasting using support vector machines”, *Neural Computing and Applications*, Vol.10(2001), 184~192.

Chawla, N., A. Lazarevic, L. Hall, and K. Bowyer, “SMOTEBoost : Improving prediction of the minority class in boosting”, 7th European conference on principles and practice of knowledge discovery in databases(2003), Cavtatv Dubrovnik, Croatia, 107~119.

Chaveesuk, R., Srivaree-Ratana, C., and A. E. Smith, “Alternative neural network approaches to corporate bond rating”, *Journal of Engineering Valuation and Cost Analysis*, Vol.2, No.2(1993), 117~131.

Cover, T. M. and J. A. Thomas, Element of information theory, John Wiley and Sons, 1991.

Darbellay, G. A., “An estimator of the mutual information based on a criterion for independence”, *Computational Statistics and Data Analysis*, Vol.32(1999), 1~17.

Dutta, S. and S. Shekhar, “Bond rating : A non-conservative application application of neural networks”, *Proceedings of IEEE International Conference on Neural Networks*, (1988), II443~II450.

Ederington, H. L., “Classification models and bond ratings”, *Financial Review*, Vol.20, No.

4(1985), 237～262.

Fawcett, T. and F. Provost, “Adaptive fraud detection”, *Data Mining and Knowledge discovery*, Vol.1, No.3(1997), 291～316.

Fisher, L., “Determinants of risk premiums on corporate bonds”, *Journal of Political Economy*, Vol.67(1959), 217～237.

Freund, Y. and R. E. Schapire, “A decision theoretic generalization of online learning and an application to boosting”, *Journal of Computer and System Science*, Vol.55, No.1(1997), 119～139.

Gentry, J. A., Whitford, D. T., and P. Newbold, “Predicting industrial bond ratings with a probit model and funds flow components”, *Financial Review*, Vol.23, No.3(1988), 269～286.

Hong, X., “A kernel-based two-class classifier for imbalanced data sets”, *IEEE Transactions on neural networks*, Vol.18, No.1(2007), 28～40.

Hsu, C. W. and C. J. Lin, “A Comparison of Methods for Multiclass Support Vector Machines”, *IEEE Transactions on Neural Networks*, Vol.13, No.2(2002), 415～425.

Huang, Zan, Chen, Hsinchun, Hsu, Chia-Jung, Chen, Wun-Hwa, and Wu, Soushan, “Credit rating analysis with support vector machines and neural networks. A market comparative study”, *Decision Support Systems*, Vol.37(2004), 543～558.

Jackson, J. D. and J. W. Boyd, “A statistical approach to modeling the behavior of bond raters”, *The Journal of Behavioral Economics*, Vol.17, No.3(1988), 173～193.

Kim, K., “Financial time series forecasting using support vector machines”, *Neurocomputing*, Vol.55(2004), 307～319.

Kotsiantis, S., D. Tzelepis, E. Kounmanakos, and V. Tampakas, “Selective costing voting for bankruptcy prediction”, *International Journal of Knowledge-based and Intelligent Engineering Systems*, Vol.11(2007), 115～127.

Kubat, M., Holte, R., and S. Matwin, “Learning when Negative example abound”, Proceedings of the 9th European Conference on Machine Learning, ECML’97, 1997.

Kwon, Y. S., Han, I. G., and K. C. Lee, “Ordinal Pairwise Partitioning (OPP) approach to neural networks training in bond rating”, *Intelligent Systems in Accounting, Finance and Management*, Vol.6(1997), 23～40.

Maher, J. J. and T. K. Sen, “Predicting bond ratings using neural networks : A comparison with logistic regression”, *Intelligent Systems in Accounting, Finance and Management*, Vol.6(1997), 59～72.

Maia, T. T., A. P. Braga, and A. F. Carvalho, “Hybrid classification algorithms based on boosting and support vector machines”, *Kybernetes*, Vol.37, No.9(2008), 1469～1491.

Min, S. H., J. M. Lee, and I. G. Han, Hybrid genetic algorithms and support vector machines for bankruptcy prediction, *Expert Systems with Applications*, Vol.31(2006), 652～660.

Optiz, D. and R. Maclin, “Popular ensemble methods : an empirical study”, *Journal of Artificial Intelligence*, Vol.11(1999), 169～198.

Pinches, G. E. and K. A. Mingom, “A multivariate analysis of industrial bond ratings”, *Journal of Finance*, Vol.28, No.1(1973), 1～18.

Platt, J., “Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf, C. Burges, and A. Smola, (Eds.)”, Advances in Kernel MethodsSupport Vector Learning, MIT Press, 1998.

Pogue, T. F. and R. M. Soldofsky, "What's in a bond rating?", *Journal of Financial and Quantitative Analysis*, Vol.4, No.2(1969), 201~228.

Seiffert, C., T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost : Improving classification performance when training data is skewed", *19th International Conference on Pattern Recognition*, (2008), 1~4.

Shin, H. J. and S. Z. Cho, "Response modeling with support vector machines", *Expert Systems with applications*, Vol.30, No.4(2006), 746~760.

Shin, K., T. Lee, and H. Kim, "An application of support vector machines in bankruptcy prediction", *Expert Systems with Applications*, Vol.28(2005), 127~135.

Tay, F. E. J. and L. J. Cao, "Modified support vector machine in financial time series forecasting", *Neurocomputing*, Vol.48(2002), 847~861.

Vapnik, V. N., "The nature of statistical learning theory", New York : Springer, 1995.

Wang, B. X. and N. Japkowicz, "Boosting support vector machines for imbalanced data sets", *Knowledge and Information Systems*, Vol.25(2010), 1~10.

Weiss, G. M., "Mining with rarity : A unifying framework", *SIGKDD Explorations*, Vol.T, No.1(2004), 7~19.

West, R. R., "An alternative approach to predicting corporate bond ratings", *Journal of Accounting Research*, Vol.8, No.1(1970), 118~125.

Wu, G. and E. Chang, "Adaptive feature-space conformal transformation for imbalanced data learning", *Proceedings of the 20th International Conference on Machine Learning*, 2003.

Wu, G. and E. Chang, "KBA : Kernel boundary alignment considering imbalanced data distribution", *IEEE Transactions on knowledge and data engineering*, Vol.17, No.6(2005), 786~795.

Wu, G. Y. Wu, L. Jiao, Y. F. Wang, and E. Chang, "Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance", *Proceedings of 20th International Conference on Multimedia*, 2003.

Abstract

Ensemble Learning with Support Vector Machines for Bond Rating

Myoung-Jong Kim*

Bond rating is regarded as an important event for measuring financial risk of companies and for determining the investment returns of investors. As a result, it has been a popular research topic for researchers to predict companies' credit ratings by applying statistical and machine learning techniques. The statistical techniques, including multiple regression, multiple discriminant analysis (MDA), logistic models (LOGIT), and probit analysis, have been traditionally used in bond rating. However, one major drawback is that it should be based on strict assumptions. Such strict assumptions include linearity, normality, independence among predictor variables and pre-existing functional forms relating the criterion variablesand the predictor variables. Those strict assumptions of traditional statistics have limited their application to the real world.

Machine learning techniques also used in bond rating prediction models include decision trees (DT), neural networks (NN), and Support Vector Machine (SVM). Especially, SVM is recognized as a new and promising classification and regression analysis method. SVM learns a separating hyperplane that can maximize the margin between two categories. SVM is simple enough to be analyzed mathematical, and leads to high performance in practical applications. SVM implements the structuralrisk minimization principle and searches to minimize an upper bound of the generalization error. In addition, the solution of SVM may be a global optimum and thus, overfitting is unlikely to occur with SVM. In addition, SVM does not require too many data sample for training since it builds prediction models by only using some representative sample near the boundaries called support vectors. A number of experimental researches have indicated that SVM has been successfully applied in a variety of pattern recognition fields.

However, there are three major drawbacks that can be potential causes for degrading SVM's performance. First, SVM is originally proposed for solving binary-class classification problems. Methods for combining SVMs for multi-class classification such as One-Against-One, One-Against-All have been proposed, but they do not improve the performance in multi-class classification problem as much as SVM for binary-class classification. Second, approximation algorithms (e.g. decomposition

* Corresponding Author: Myoung-Jong Kim
School of Business, Pusan National University San 30, Jangjeon, Geumjeong, Busan 609-735, Korea
Tel: +82-51-510-3154, Fax: +82-51-581-3144, E-mail: mjongkim@pusan.ac.kr

methods, sequential minimal optimization algorithm) could be used for effective multi-class computation to reduce computation time, but it could deteriorate classification performance. Third, the difficulty in multi-class prediction problems is in data imbalance problem that can occur when the number of instances in one class greatly outnumbers the number of instances in the other class. Such data sets often cause a default classifier to be built due to skewed boundary and thus the reduction in the classification accuracy of such a classifier.

SVM ensemble learning is one of machine learning methods to cope with the above drawbacks. Ensemble learning is a method for improving the performance of classification and prediction algorithms. AdaBoost is one of the widely used ensemble learning techniques. It constructs a composite classifier by sequentially training classifiers while increasing weight on the misclassified observations through iterations. The observations that are incorrectly predicted by previous classifiers are chosen more often than examples that are correctly predicted. Thus Boosting attempts to produce new classifiers that are better able to predict examples for which the current ensemble's performance is poor. In this way, it can reinforce the training of the misclassified observations of the minority class.

This paper proposes a multiclass Geometric Mean-based Boosting (MGM-Boost) to resolve multiclass prediction problem. Since MGM-Boost introduces the notion of geometric mean into AdaBoost, it can perform learning process considering the geometric mean-based accuracy and errors of multiclass. This study applies MGM-Boost to the real-world bond rating case for Korean companies to examine the feasibility of MGM-Boost. 10-fold cross validations for threetimes with different random seeds are performed in order to ensure that the comparison among three different classifiers does not happen by chance. For each of 10-fold cross validation, the entire data set is first partitioned into tenequal-sized sets, and then each set is in turn used as the test set while the classifier trains on the other nine sets. That is, cross-validated folds have been tested independently of each algorithm. Through these steps, we have obtained the results for classifiers on each of the 30 experiments.

In the comparison of arithmetic mean-based prediction accuracy between individual classifiers, MGM-Boost (52.95%) shows higher prediction accuracy than both AdaBoost (51.69%) and SVM (49.47%). MGM-Boost (28.12%) also shows the higher prediction accuracy than AdaBoost (24.65%) and SVM (15.42%)in terms of geometric mean-based prediction accuracy. T-test is used to examine whether the performance of each classifiers for 30 folds is significantly different. The results indicate that performance of MGM-Boost is significantly different from AdaBoost and SVM classifiers at 1% level. These results mean that MGM-Boost can provide robust and stable solutions to multi-class-problems such as bond rating.

Key Words : Support Vector Machine, AdaBoost, Multiclass Geometric Mean-based Boosting, Bond Rating Prediction

저 자 소 개

김명종

성균관대학교 회계학과, 동대학원에서 경영학 석사학위 취득 후 한국과학기술원에서 경영공학박사 학위를 취득하였다. 현재 부산대학교에 경영학과 교수로 재직하고 있으며 주요 관심분야는 회계, 재무, 데이터마이닝, 지식공학 등의 결합 메커니즘이다.