지능정보연구 제15권 제3호 2009년 9월(pp.01~15)

기업부실 예측 데이터의 불균형 문제 해결을 위한 앙상블 학습

김명종
동서대학교 경영학부
(mjongkim@gdsu.dongseo.ac.kr)

데이터 불균형 문제는 분류 및 예측 문제에서 하나의 범주에 속하는 표본의 수가 다른 범주들에 속하는 표본 수에 비하여 현저하게 적을 경우 나타난다. 데이터 불균형이 심화됨에 따라 범주 사이의 분류경계영역이 왜곡되고 결과적으로 분류자의 학습성과가 저하되는 문제가 발생한다. 본 연구에서는 데이터 불균형 문제를 해결하기 위하여 Geometric Mean-based Boosting (GM-Boost) 알고리즘을 제안하고자 한다. GM-Boost 알고리즘은 기하평균 개념에 기초하고 있어 다수 범주와 소수 범주를 동시에 고려한 학습이 가능하고 오분류된 표본에 집중하여 학습을 강화할 수 있는 장점이 있다. 기업부실 예측문제를 활용하여 GM-Boost 알고리즘의 성과를 검증한 결과 기존의 Under-Sampling, Over-Sampling 및 AdaBoost 알고리즘에 비하여 우수한 분류 정확성을 보여주었고 데이터 불균형 정도에 관계없이 견고한 학습성과를 나타냈다.

논문접수일 : 2009년 05월 21일 논문수정일 : 2009년 06월 30일 게재확정일 : 2009년 07월 01일 교신저자 : 김명종

1. 서론

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; Yan et al., 2003) 등의 분류 및 예측 문제에서 빈번히 관찰되는 데이터 불균형(Data imbalance) 은 사용되는 표본이 하나의 범주에 편중되었을 때 나타난다.

데이터 불균형에서 파생되는 문제점은 다음과 같다(강필성 외, 2006; Kotsiantis et al., 2007; Wang and Japkowicz, 2009). 첫째, 성과지표의 적합성 문제이다. 현재까지 분류자의 성과 측정에 보편적으로 활용되는 지표는 단순평균 정확도로 전체 표본 중 정분류된 표본의 비율로 계산된다. 그러나 데이터 불균형이 존재하는 상황에서 단순평균 정확도는 다수 범주(majority class) 표본의 분류 정확성에 의존하여 분류자의 성과를 결정하는 단점이 있다. 예를 들어 기업의 부실은 발생 빈도가 매우 희귀한 사건으로 무디스(Moody's) 등 전문 신용평가기관은 국내 외부감사 법인의 장기평균 부도율을 약 3~5% 수준으로 예상하고 있다. 만일 전체 외부감사 기업을 학습 자료로 활용하는 경우, 단순평균 정확도는 다수 범주인 정상 기업의 분류 정확성에 의존하여 지속적으로 증가하지만, 정작 부실기업에 대한 분류 정확성이 감소하게 된다. 이러한 문제로 인하여 최근 다수 범주와 소수 범주의

정확도를 동시에 고려할 수 있는 ROC 분석(Receiving Operator Characteristic Analysis)이나 기하평균 정확도(Geometric-Mean Accuracy)와 같은 지표들이 단순평균 정확도를 대체하여 이용되고 있다(Kubat et al., 1997; Fawcett, 2006).

둘째, 분류자의 학습 성과가 저하되는 문제이다. 데이터 불균형 하에서 다수 범주 표본에 의한 분류경계영역의 침해로 인하여 소수 범주 영역이 점차로 축소하고 결과적으로 소수 범주에 대한 분류 정확성이 급격히 감소된다. 이러한 문제의 해결 대안으로서 Under-Sampling, Over-Sampling, Cost Adaptation Strategies, 부스팅 알고리즘 등 다양한 기법이 활용되고 있다. Under-Sampling 기법은 정해진 규칙에 의해 소수 범주의 표본 수와 동일하게 다수 범주의 표본을 추출하는 방법이다. Over-Sampling 방법은 Under-Sampling 방법과 정반대의 방법으로 정해진 규칙에 의해 다수 범주의 표본 수만큼 소수 범주의 표본을 증가시키는 방법이다. Cost-Adaptation Strategies는 오분류 관측치에 패널티를 부과하는 방식으로 데이터 분포를 왜곡시키지 않는다는 장점이 있는 반면 데이터 불균형이 매우 심각할 경우 효과가 미미하다는 단점이 있다. 최근 SMOTEBoost (Chawla et al., 2003), RUSBoost(Seiffert et al., 2008) 및 AdaBoost 등 다양한 부스팅 알고리즘 (Boosting algorithms)이 데이터 불균형 문제의 해결 대안으로 제안되고 있다. 이러한 부스팅 알고리즘은 소수 범주 표본에 대한 학습을 효과적으로 진행할 수 있다는 장점을 가지고 있다.

기업부실 예측연구에서 보편적으로 이용되고 있는 샘플링 방법은 Under-Sampling으로서 학습시간이 단축된다는 장점이 있다. 그러나 추출 표본이 모집단의 특성을 대표하지 못하는 경우 분류자의 일반화 특성이 감소되며, 소수 범주의 표본 수가 매우 작은 경우 분류 정확성이 심각하게 훼손될 위험이 있다. 이와 반대로 실무에서는 별도의 샘플링이 없이 전체 기업 표본을 이용하여 기업부실 예측모형을 개발하는 관행이 간혹 관찰된다. 그러나 부실기업 수가 매우 적다는 점을 고려할 때 이러한 개발 관행은 분류자의 정확성 변화에 대한 민감도 분석이 필수적으로 병행되어야 한다.

본 연구에서는 데이터 불균형의 문제점을 완화하고 다수 범주와 소수 범주에 대한 균형적 학습이 가능한 Geometric Mean-based Boosting (GM-Boost)을 제안하고자 한다. GM-Boost는 AdaBoost 알고리즘에 Kubat et al.(1997)이 제안한 기하평균 개념을 도입한 새로운 부스팅 알고리즘으로 데이터 불균형이 심각한 상황에서도 높은 예측력을 확보할 수 있으며 데이터 불균형 정도에 관계없이 견고한 학습능력을 제공한다는 장점이 있다. 본 연구에서는 GM-Boost를 데이터 불균형이 존재하는 2범주 분류 문제인 기업부실 예측에 적용하여 제안 알고리즘의 성과를 검증하고자 한다.

본 연구는 다음과 같이 구성되어 있다. 제 2장에서는 데이터 불균형의 문제점 및 이를 해결하기 위한 기존의 방법을 고찰하고자 한다. 제 3장에서는 AdaBoost 알고리즘과 GM-Boost 알고리즘에 대하여 비교 설명하고자 한다. 제 4장에서는 제안모형의 유용성을 확인하기 위한 실험 데이터 수집 및 실험 설계 과정에 대하여 설명한다. 제 5장에서는 GM-Boost의 성과 검증 결과를 종합적으로 정리하여 제시하고자 한다. 마지막 제 6장에서는 결론과 함께 향후 연구방향을 제시하고자 한다.

2. 기업부실 예측문제의 데이터 불균형

본 장에서는 데이터 불균형의 문제를 살펴보고 이를 해결하기 위하여 제안된 해결방법을 검토하고자 한다.

2.1 성과 지표의 유효성

정확도의 개념을 설명하기 위하여 소수 범주에 속하는 표본을 Positive라고 하고 다수 범주에 속하는 표본을 Negative라고 하면 <표 1>과 같은 오분류표(Confusion matrix)를 도출할 수 있다. 현재 보편적으로 활용되고 있는 단순평균 정확도는 (TP + TN)/(TP + FN + FP + TN)으로 계산된다. 이미 언급한 바와 같이 단순평균 정확도는 데이터 균형 하에서는 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; Kotsiantis et al., 2007; Wang and Japkowicz, 2009).

<표 1> 오분류표

		예측 범주	
		Positive	Negative
실제 범주	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

이러한 문제점에 대한 해결대안으로 제안된 방법이 기하평균 정확도와 ROC 분석이다. 기하평균 정확도는 다수 범주의 정확도와 소수 범주의 정확도를 모두 고려한 성과지표로 $(\text{민감도} \times \text{특이도})^{1/2}$로 계산된다(Kubat et al., 1997). 여기에서 민감도와 특이도는 각각 TP/(TP + FN)과 TN/(FP + TN)으로 측정된다.

ROC 분석은 분류자 결과 값의 서열화 순위에 따라 수평축에는 1-특이도, 수직 축에는 민감도를 표시하여 연결한 ROC 곡선에 기초하여 분류자의 정확성을 분석하는 방법이다. 분류자의 정확도는 ROC 곡선의 면적(Area Under ROC : AUROC)로 계산된다. 완벽한 모형의 AUROC는 1이 되며, 임의추측 모형의 AUROC는 0.5가 된다. 대부분의 모형은 일반적으로 0.5보다 크고 1보다 작은 AUROC를 가지며 AUROC가 1에 근접할수록 정확도가 높은 모형으로 평가된다(Fawcett, 2006).

2.2 기업부실 예측연구에서의 데이터 불균형 문제

기존에 기업부실 예측모형의 개발을 위하여 활용되었던 기법은 크게 통계학습과 기계학습으로 구분된다. 통계학습 분야에서는 Beaver(1966)의 단변량 분석을 효시로 다중 회귀분석(Meyer and Pifer, 1970), 판별분석(Altman, 1968; Altman et al., 1977), 로지스틱 회귀분석 (Dimitras et al., 1996; Ohlson, 1980; Pantalone and Platt, 1987), 프로빗 모형(Zmijewski, 1984) 등 다양한 통계적 기법들이 부실예측 모형에 활용되어 왔다. 기계학습 분야에서는 의사결정트리, 인공신경망, 사례기반추론, SVM 등 다양한 기법들이 기업부실 예측에 적용되었다 (Han et al., 1996; Shaw and Gentry, 1998; Bryant, 1997; Buta, 1994; Odom and Sharda, 1990; Laitinen and Kankaanpaa 1997; Ravi and Ravi, 2007).

특히, 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. 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 원칙(structural risk minimization)에 기반하므로 과대적합 (overfitting) 문제에 견고하다는 장점이 있다. 이러한 장점

으로 인하여 문자인식, 이미지 인식, 마이크로어레이 분석 등 자연과학 분야에서 적용되어 왔으며 최근 시계열 예측 및 분류(Cao and Tay, 2001; Kim, 2004; Tay and Cao, 2002), 채권신용등급(Huang et al., 2004), 기업부실예측(Shin et al., 2005; Min et al., 2006) 등 경영분야에도 활발하게 적용되고 있다.

SVM을 기업부실 예측에 활용한 연구로서 Shin et al.(2005)은 정상기업과 부실기업의 구성비율이 1 : 1인 2320개 기업을 대상으로 SVM과 인공신경망의 분류 정확성을 비교하였다. 실험 결과 SVM은 인공신경망보다 우수한 분류 정확성을 보여주었다. Min et al.(2006)은 각 307개의 정상기업과 부실기업을 대상으로 SVM과 유전자 알고리즘의 결합분류자, SVM, 인공신경망, 로지스틱 회귀분석의 분류 정확성을 비교하였다. 실험 결과 SVM 기반의 분류자들은 로지스틱 회귀분석이나 인공신경망보다 높은 분류 정확성을 보여주었다. 이와 같이 대부분의 기업부실 예측연구들은 Under-Sampling 방법을 이용하여 데이터 불균형 문제를 통제하고 있다. 그러나 이러한 접근방법은 학습 표본의 모집단 대표성 여부에 따라 분류자의 일반화 성능이 크게 달라질 수 있다는 단점이 있다. 또한 다수 범주 표본의 분포를 왜곡시키고 소수 범주에 비정상적으로 초점을 맞춘 학습이라는 비판이 제기되고 있다(강필성 외, 2006).

이와는 반대로, 일부 금융기관에서는 표본에 대한 샘플링 절차를 생략하고 전체 기업 모집단을 이용하여 기업부실 예측모형을 개발하고 있다. 이러한 접근방법은 데이터 불균형으로 인하여 부실기업의 분류 정확도가 낮아질 수 있다는 문제점이 있다. 특히 기업의 부실화는 기업의 이해관계자와 더불어 국가경제에 막대한 손실을 초래하므로 이러한 접근방법의 타당성 검증이 수행되어야 한다.

강필성 외(2006)의 연구에서는 2범주 분류 문제에서 데이터 불균형이 SVM의 분류 정확성에 미치는 영향을 분석하기 위하여 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 표본집합(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 데이터 불균형 문제의 해결 방법

기존 연구에서 데이터 불균형 문제를 해결하기 위하여 제안된 방법은 크게 데이터 중심의 방법과 분류자 중심의 방법으로 구분된다(Kotsiantis et al., 2007).

데이터 중심의 접근 방법 중 보편적으로 활용되고 있는 Under-Sampling은 소수 범주 표본에 맞추어 다수 범주의 표본을 무작위 또는 특정 규칙을 이용하여 제거하는 방법이다. 이 방법은 다수 범주에 대한 정보손실이 필수적으로 발생한다는 단점이 있다. 하지만 효과적인 규칙을 이용하는 경우, 데이터 불균형 문제를 효과적으로 해소하였음을 보여주었다(Laurikkala, 2002; Japkowicz and Stephen, 2002; kubat and Matwin, 1997). Over-Sampling은 데이터 복제 및 데이터 생성 등의 방법을 이용하여 소수 범주 표본을 증가시키는 방법이다(Chawla et al., 2002; Japkowicz and Stephen, 2002). 이 방법은 다수 범주 표본의 정보손실이 없이 표본을 증가시킬 수 있다는 장점이 있는 반면, 데이터 수의 증가에 따라 학습 소요시간이 증가한다는 단점이 있다. 많은 연구에서 Under-Sampling이 Over-Sampling에 비하여 데이터 불균형 문제를 효과적으로 해결하는 것으로 보고되고 있다. 이는 Over-Sampling의 데이터 생성과정에서 소수 범주 데이터의 편차가 크게 증가하기 때문이다(Wang and Japkowicz, 2009).

분류자 중심의 방법은 일반적으로 Cost-adaptation learning strategies를 이용하여 오분류된 패턴에 서로 다른 패널티를 부과한다. 즉 소수 범주에 속하는 데이터가 오분류된 경우, 다수 범주에 속하는 데이터가 오분류되는 경우보다 높은 패널티를 부과하는 방법이다(Elkan, 2001; Provost and Fawcett, 2001). 이 방법은 Under-Sampling방법의 정보손실이나 Over-Sampling 방법의 편차증가와 같은 문제는 발생하지 않지만 표본에 대한 과도한 민감성으로 인하여 불안정한 분류자가 생성될 수 있다는 단점이 있다.

최근 SMOTEBoost(Chawla et al., 2003), RUSBoost (Seiffert et al., 2008) 등의 부스팅 알고리즘이 데이터 불균형 문제에 적용되어 성공적인 결과를 보여주었다. 부스팅 알고리즘은 오분류된 관측치에 초점을 맞추어 분류자 앙상블을 순차적으로 생성하는 방법으로 오분류율이 높은 소수 범주 표본에 대한 학습을 강화할 수 있다는 장점이 있다.

3. GM-Boost 알고리즘

본 장에서는 대표적 부스팅 기법인 AdaBoost 알고리즘과 본 연구에서 제안한 GM-Boost 알고리즘에 대하여 비교하여 설명하고자 한다.

3.1 AdaBoost 알고리즘

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 순차적으로 생성한다. AdaBoost 알고리즘의 설명을 위하여 n개의 학습 표본과 K개의 기저 분류자로 구성된 앙상블 $C = \{C_1, C_2, \cdots, C_K\}$을 가정하면 k번째 기저분류자의 오류율(e_k)은 다음과 같이 단순평균으로 계산된다.

$$e_k = \frac{1}{n}\sum_{i=1}^{n} L(C_k(x_i),\ y_i)$$

$$L(C_k(x_i),\ y_i) = \begin{cases} 1 & C_k(x_i) \neq y_i \\ 0 & C_k(x_i) = y_i \end{cases}$$

여기에서 x_i는 i번째 관측치의 예측변수 벡터이고 y_i는 i번째 관측치의 범주를 나타내며 $C_k(x_i)$는 예측변수 벡터 x_i에 대한 k번째 분류자의 분류결과이다. $k+1$번째 분류자에서 i번째 관측치에 부여되는 가중치는 $w_{k+1}(i) = w_k(i)\exp(\alpha_k L(C_k(x_i),\ y_i))$로 조정되어 오분류된 관측치에 더 높은 가중치가 부여된다. 여기에서 α_k는 분류자의 중요도 또는 정확도의 개념으로 해석되며 $\alpha_k = 1\mathrm{n}((1-e_k)/e_k)$로 계산된다. $k+1$번째 분류자의 학습표본을 구성할 때 가중치가 높은 오분류 관측치가 많이 포함되기 때문에 부스팅 알고리즘은 오분류 관측치에 초점을 맞춘 학습을 진행할 수 있게 된다. 앙상블 학습은 $e_k < 0.5$일 때 학습을 중단하며 i번째 관측치의 최종결과를 다음과 같이 앙상블의 결과치의 가중평균으로 계산하여 산출한다.

$$C(x_i) = sign\left(\sum_{k=1}^{K} \alpha_k C_k(x_i)\right)$$

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 나타난다. 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, 학습의 조기중단 문제이다. 분류자의 오류율 e_k는 단순평균 오류율로 전체 표본 대비 오분류 표본 비율로 계산된다. 불균형 데이터의 경우 다수 범주의 낮은 오류율로 인하여 단순평균 오류율은 빠르게 중단조건 ($e_k < 0.5$)에 수렴하게 되고 충분한 학습이 이루어지기도 전에 학습이 중단될 수도 있다. 둘째, 분류자의 성과를 나타내는 α_k 역시 단순평균 정확도에 기초한 개념이다. 이미 언급한 바와 같이 데이터 불균형 하에서 단순평균 정확도는 성과지표로 유효하지 않기 때문에 다수 범주와 소수 범주를 동시에 고려한 가중평균 정확도 개념으로 대체할 필요가 있다.

3.2 GM-Boost 알고리즘

앞서 언급한 AdaBoost 알고리즘 의 기본 가정과 더불어 n개의 학습 표본이 소수 범주에 속하는 n^+개의 표본과 다수 범주에 속하는 n^-개의 표본으로 구성되어 있다고 가정해보자. 이 때 k번째 분류자의 소수 범주 오류율을 e_k^+라 하고 다수 범주 오류율을 e_k^-라 하면 기하평균 오류율(e_k)은 다음과 같이 계산된다.

$$e_k = \sqrt{e_k^+ \cdot e_k^-},$$

$$\text{단, } e_k^+ = \frac{1}{n^+}\sum_{i=1}^{n^+} L(C_k(x_i),\ y_i)$$

$$\text{and } e_k^- = \frac{1}{n^-}\sum_{i=1}^{n^-} L(C_k(x_i),\ y_i)$$

이에 따라 분류자의 분류 정확성을 의미하는 α_k는 소수 범주에 대한 분류 정확도와 다수 범주에 대한 분류 정확도의 가중평균 정확도로 계산된다.

$$\alpha_k = \sqrt{\alpha_k^+ \cdot \alpha_k^-} \text{ 단, } \alpha_k^+ = \ln\left(\frac{1-e_k^+}{e_k}\right)$$

$$\text{and } \alpha_k^- = \ln\left(\frac{1-e_k^-}{e_k}\right)$$

$k+1$번째 분류자에서 표본에 부여되는 가중치는 $w_{k+1}(i) = w_k(i)\exp(\alpha_k L(C_k(x_i), y_i))$로 계산되며 i번째 관측치의 최종결과는 앙상블 결과와 α_k의 선형결합으로 산출된다.

$$C(x_i) = sign\left(\sum_{k=1}^{K}\alpha_k C_k(x_i)\right)$$

GM-Boost 알고리즘의 GM-Boost 알고리즘은 AdaBoost 알고리즘에 기초하고 있으므로 소수 범주 표본에 대한 학습기회를 강화할 수 있다는 장점을 가지고 있다. 또한 가중평균 오류율과 가중평균 정확도에 기초하므로 학습의 조기중단 문제를 해결할 수 있으며 다수 범주와 소수 범주를 동시에 고려한 학습을 진행할 수 있다는 장점이 있다. <그림 1>은 GM-Boost 알고리즘의 절차를 간략하게 요약하고 있다.

4. 연구 설계

본 연구의 실험 데이터는 한신평정보㈜의 기업정보 DB를 기초로 수집하였다. 부실 기업은 2002~2005년 중 은행연합회 신용정보 등록기업, 당좌부도 발생기업, 회사정리절차 개시기업, 기업구조조정절차 개시기업에 해당하는 400개 외부감사 제조기업으로 구성하였으며 부실기업에 대한 재무자료는 부실 직전 년도를 중심으로 수집하되 직전 년도 재무자료가 없는 경우 2년 전 재무자료를 수집하였다. 정상 기업은 2002~2005년 말 기준 부실 사유에 해당하지 않는 2,400개 외부감사 제조기업으로 2001~2004년의 년도 별 재무제표 9,600건을 수집하였다. 이러한 방법으로 재무자료 기준으로 총 10,000건의 재무 자료를 수집하였으며 4년의 평균부도율은 전문 신용평가기관의 부도율 예상 범위 (3~5%)인 4% 수준으로 유지하였다.

부실 예측에 사용되는 재무비율은 일차적으로

1. 가중치 초기화 : $w_k(x_i) = 1/n,\ i = 1, 2, \cdots, n$
2. 중단조건($e_k < 0.5$)을 만족할 때까지 분류자 생성 반복(C_k, $k = 1, 2, \cdots, K$)
   a) 가중치 기준에 의거하여 추출된 학습 데이터의 분류자 학습 : $C_k(x_i) \in \{-1, 1\}$
   b) 기하평균 오류율 산출
   $e_k = \sqrt{e_k^+ \cdot e_k^-}$
   $e_k^+ = \frac{1}{n^+}\sum_{i=1}^{n^+} L(C_k(x_i), y_i)$ and $e_k^- = \frac{1}{n^-}\sum_{i=1}^{n^-} L(C_k(x_i), y_i)$
   c) 기하평균 정확도 산출
   $\alpha_k = \sqrt{\alpha_k^+ \cdot \alpha_k^-}$ 단, $\alpha_k^+ = \ln\left(\frac{1-e_k^+}{e_k}\right)$ and $\alpha_k^- = \ln\left(\frac{1-e_k^-}{e_k}\right)$
   d) 가중치 조정 : $w_{k+1}(i) = w_k(i)\exp(\alpha_k L(C_k(x_i), y_i))$
3. 최종 결과 생성 : $C(x_i) = sign\left(\sum_{k=1}^{K}\alpha_k C_k(x_i)\right)$

<그림 1> GM-Boost 알고리즘의 개요

<표 2> 재무비율의 AUROC

분류군	재무비율	정확도	분류군	재무비율	정확도
수익성	총자산경상이익율*	51.7	레버리지	자기자본비율*	50.9
	총자산순이익율	44.7		유동자산/총자산	50.3
	금융비용/매출액	49.0	자본구조	이익잉여금/총자산*	52.5
	금융비용/총부채	47.3		이익잉여금/총부채	51.7
	순금융비용/매출액	49.2		이익잉여금/유동자산	50.1
	매출액경상이익율	44.5	유동성	현금비율*	45.5
	매출액순이익율	49.1		당좌비율	45.1
	자기자본경상이익율	47.2		유동비율	42.2
	자기자본순이익율	47.1	활동성	재고자산회전율*	30.5
부채상환	EBITDA/이자비용*	51.2		유동부채회전율	28.3
	EBIT/이자비용	48.2		매출채권회전율	27.2
	영업현금흐름/이자비용	47.5	규모	총자산*	24.2
	영업현금흐름/총부채	47.1		매출액	21.4
	잉여현금흐름/이자비용	50.5		고정자산	22.6
	잉여현금흐름/총부채	50.1			
	부채상환계수	48.8			

주) * 최종 7개 재무비율.

기존의 기업부실 예측연구에 사용된 비율 및 실무에서 부실예측의 지표로 유용하게 활용되는 비율을 중심으로 30개의 재무비율을 수집하였다. 수집된 재무비율을 수익성, 부채상환능력, 레버지리, 자본구조, 유동성, 활동성 및 규모의 7개 재무비율군으로 재분류하였으며, 최종 입력변수는 데이터 불균형을 고려하여 ROC 곡선에 의하여 산출된 AUROC를 이용하여 각 분류군별로 AUROC가 높은 7개 재무비율을 선정하였다. 일차적으로 선정된 30개 비율 및 최종 선정된 7개 재무비율의 AUROC는 <표 2>에 제시되어있다.

비록 변별력과 직접적인 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. 일반적으로 다중공선성이 존재한다고 의심되는 VIF 임계치는 5~10사이이며 VIF가 10이상이면 다중공선성이 심각한 것으로 판단할 수 있다. 따라서 <표 3>은 최종 선정된 7개 재무비율 간에는 다중공선성이 실질적으로 존재하지 않음을 나타내고 있다.

<표 3> 분산팽창요인 분석의 결과

재무비율	VIF
총자산경상이익율	1.36
EBITDA/이자비용	2.11
자기자본비율	1.77
이익잉여금/총자산	2.53
현금비율	1.34
재고자산회전율	1.59
총자산	1.31

5. 연구 결과

본 장에서는 GM-Boost의 성과 검증 결과를 종합적으로 정리하여 제시하고자 한다.

GM-Boost 알고리즘의 성과비교 대상으로 Under-Sampling, Over-Sampling, AdaBoost 알고리즘을 선정하였다. 기저 분류자(base classifier)로서 SVM

<표 4> 데이터 구성현황

표본집합		학습표본			검증표본		
집합	비율	부실	정상	총계	부실	정상	총계
A	1 : 1	300	300	600	100	100	200
B	1 : 3	300	900	1,200	100	300	400
C	1 : 5	300	1,500	1,800	100	500	600
D	1 : 10	300	3,000	3,300	100	1,000	1,100
E	1 : 24	300	7,200	7,500	100	2,400	2,500

은 Platt(1998)에 의해 제안된 SMO(Sequential Minimal Optimization) 알고리즘을 사용하였다. 데이터를 고차원 특성공간으로 매핑하기 위한 커널함수는 커널차수가 1인 다항커널 (polynomial kernel)을 이용하였다. AdaBoost 알고리즘과 GM-Boost 알고리즘은 분류자 생성횟수가 25회를 넘어서면 오류 감소효과가 미미하다는 연구 결과(Opitz and Maclin, 1999)에 기초하여 최대 앙상블 생성횟수를 25회로 제한하였다.

표본 불균형에 따른 분류자의 성과 변화를 분석하기 위하여 표본 구성은 다음과 같이 2단계로 나누어 진행하였다. 1단계에서는 전체 10,000개 표본 중에서 부실기업과 정상 기업을 1 : 1(A), 1 : 3(B), 1 : 5(C), 1 : 10(D), 1 : 24(E)의 비율로 5개 표본을 추출하였고 각 표본의 75%는 학습표본으로 25%는 검증표본으로 구분하였다. 이러한 방법으로 추출된 데이터 구성현황은 <표 4>에 요약되어 있다.

같은 방법을 50회 반복하여 각 표본 집합마다 50개의 학습표본과 각 50개의 검증표본을 구성하였다.

2단계에서는 1단계에서 구성된 표본을 다음과 같이 재구성하였다. Under-Sampling 방법을 적용하기 위하여 표본집합 B, C, D, E의 학습표본에서 부실기업 수(n = 300)와 동일하게 정상 기업을 무작위로 추출하여 부실기업과 정상 기업을 1 : 1비율로 재구성하였다. Over-Sampling의 학습표본은 Chawla(2002)가 제안한 SMOTE(Synthetic Minority Over-Sampling Technique) 알고리즘을 이용하여 새로운 부실기업 데이터를 생성하였다. SMOTE 알고리즘은 k minority class nearest neighbors에 기초하여 특정 관측치 및 유사한 k개의 소수범주의 사례를 조합하여 새로운 데이터를 생성하는 방법으로 일반적으로 $X_{\neq w} = X + raand(0, 1) \times (X_{nn} - X)$를 이용하여 새로운 사례를 생성한다. 여기에서 X_{new}, X, X_{nn}는 각각 새로운 생성 데이터, 표본 데이터 및 가장 유사한 k개의 사례를 의미하며 본 연구에서는 k를 1로 설정하였다. 이 방법은 표본데이터와 가장 유사한 데이터를 추출하고 표본 데이터와 유사 데이터의 차이(거리)에 임의의 난수를 곱하여 구한 값을 표본 데이터에 가산하는 방식으로 새로운 데이터를 생성하게 된다. 이러한 방법으로 부실기업과 정상기업의 비율이 1 : n인 표본에 대하여 n-1번을 반복하여 정상기업과 동일한 부실기업수를 생성하게 된다. AdaBoost 알고리즘과 GM-Boost 알고리즘은 1단계에서 구성된 표본 집합을 별도의 처리 없이 활용하였다.

최종 구성된 표본 집합을 대상으로 50회 교차타당성 검증을 수행하였으며 Duncan test를 이용하여 GM-Boost와 타 분류자의 분류 정확도 사이에 정확성 차이를 분석하였다. <표 5>는 5개 표본집합에 대한 기하평균 정확도의 50회 평균과 Duncan

test 결과를 제시하고 있다.

<표 5> 불균형 데이터에 대한 각 분류자의 분류 정확도(기하평균 정확도) 평균

표본 집합	SVM	OS-SVM	US-SVM	Ada Boost	GM-Boost
A	0.7297*	0.7297*	0.7297*	0.7399	0.7411
B	0.7255*	0.7332*	0.7347*	0.7447*	0.7524
C	0.7058*	0.7318*	0.7347*	0.7495*	0.7607
D	0.6833*	0.7233*	0.7263*	0.7422*	0.7620
E	0.6809*	0.7247*	0.7240*	0.7417*	0.7617

주) * 1% 수준에서 유의, ** 5%수준에서 유의.

<표 6>은 데이터 불균형이 가장 심한 표본집합 E에 대한 각 분류자의 분류 정확도 사이에 통계적으로 유의한 차이가 있는지를 검토한 T-test 분석 결과를 보여주고 있다.

<표 6> 표본집합 E에 대한 T-test 결과

분류자	OS-SVM	US-SVM	AdaBoost	GM-Boost
SVM	15.4670*	15.2367*	21.4735*	28.5474*
OS-SVM	-	-0.2303	6.0065*	13.0804*
US-SVM	-	-	6.2368*	13.3107*
AdaBoost	-	-	-	7.0739*

주) * 1% 수준에서 유의, ** 5%수준에서 유의.

<표 5>와 <표 6>은 다음과 같은 결과들을 제시하고 있다. 첫째, US-SVM이나 OS-SVM은 샘플링을 수행하지 않은 SVM에 비하여 모든 표본집합에서 유의적인 분류 정확성 차이를 보여주고 있다. 특히, 데이터 불균형이 심할수록 분류 정확도 차이가 더욱 커지고 있으며 표본집합 E에 대한 T-test 결과 1% 유의수준 하에서 유의적인 성과차이를 보이고 있다. 이는 Sampling을 이용하는 방법이 Sampling을 하지 않는 방법보다 데이터 불균형 해소에 효과적임을 의미한다.

둘째, 부스팅 알고리즘 기반의 AdaBoost와 GM-Boost 알고리즘은 US-SVM과 OS-SVM과 비교하여 높은 분류 성과를 나타내고 있으며 불균형이 심할수록 성과 격차도 점점 커지고 있다. 표본집합 E에 대한 T-test 결과도 1% 유의수준에서 두 방법과 비교하여 유의적인 성과차이를 보여주고 있다.

셋째, GM-Boost 알고리즘은 AdaBoost 알고리즘과 비교하여 데이터 균형인 표본집합 A를 제외한 모든 불균형 표본집합에서 유의적인 분류 정확성 차이를 보여주고 있다. 표본집합 E에 대한 T-test 결과 역시 유의적인 분류 정확도 차이를 보여주고 있다. 결과적으로 소수 범주와 다수 범주에 대한 균형적 학습이 가능한 GM-Boost 알고리즘은 AdaBoost 알고리즘과 비교하여 데이터 불균형 문제에 대한 효과적인 해결 대안이 될 수 있다.

넷째, GM-Boost 알고리즘은 모든 표본 집합에 대하여 가장 높은 분류 정확성을 보여주었다. <표 6>의 Duncan test 결과에서도 GM-Boost는 AdaBoost의 표본집합 A를 제외하고 다른 분류자에 비교하여 유의적인 정확도 차이를 보여 주었다. 이러한 결과는 GM-Boost 알고리즘이 데이터 불균형에 관계없이 높은 정확성과 견고한 학습능력을 보유하고 있음을 의미한다.

6. 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 GM-Boost 알고리즘을 제안하였다.

기업 부실 예측문제를 대상으로 GM-Boost 알고리즘의 성과를 확인한 결과 GM-Boost 알고리즘은 데

이터 불균형이 심각한 상황에서도 높은 분류 정확성과 견고한 학습능력을 확보하고 있음을 확인하였다.

본 연구의 한계를 해결하기 활용하기 위하여 다음과 같은 후속연구가 진행되기를 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).

셋째, 데이터 불균형이 발생하는 영역 중 GM-Boost 알고리즘이 적용 가능한 또 다른 영역으로는 다범주 분류문제를 들 수 있다. 특히 회사채 평가 등은 재무분야의 전형적인 다범주 분류문제라 할 수 있으며 후속연구로서 이 영역에 대한 연구를 수행하고자 한다.

참고문헌

강필성, 조성준 (2006), "데이터 불균형 해결을 위한 Under-sampling 기반 앙상블 SVMs", *대한산업공학회/한국경영과학회 2006 춘계공동학술대회*.

Altman, E. L., "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The Journal of Finance*, Vol.23 No.4 (1968), 589~609.

Altman, E. L., I. Edward, R. Haldeman, and P. Narayanan, "A new model to identify bankruptcy risk of corporations", *Journal of Banking and Finance*, Vol.1(1977), 29~54.

Beaver, W., "Financial ratios as predictors of failure, empirical research in accounting : Selected studied", *Journal of Accounting Research*, Vol.4, No.3(1966), 71~111.

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.

Bryant, S. M., "A case-based reasoning approach to bankruptcy prediction modeling", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.6, No.3(1997), 195~214

Buta, P., "Mining for financial knowledge with CBR", *AI Expert*, Vol.9. No.10(1994), 34~41.

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

Chawla, N., K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE : synthetic minority oversampling techniques", *Journal of Artificial Intelligence Research*, Vol.16(2002), 321~357.

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*. Cavtat-Dubrovnik, Croatia, (2003), 107~119.

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 inde-

pendence", *Computational Statistics and Data Analysis*, Vol.32(1999), 1~17.

Dimitras, A. I., S. H. Zanakis, and C. Zopounidis, "A survey of business failure with an emphasis on prediction methods and industrial applications", *European Journal of Operational Research*, Vol.90, No.3(1996), 487~513.

Elkan, C., "The foundation of cost-sensitive learning", In P*roceedings of the 17th International Joint Conference on Artificial Intelligence*, (2001), 973~978, Seattle, WA.

Fawcett, T., "An introduction to ROC analysis", *Pattern Recognition Letters*, Vol.27(2006), 861~874.

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

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.

Han, I., J. S. Chandler, and T. P. Liang, "The impact of measurement scale and correlation structure on classification performance of inductive learning and statistical methods". *Expert System with Applications*, Vol.10, No.2(1996), 209~221.

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

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.

Japkowicz, N. and S. Stephen, "The class imbalance problem : a systematic study", *Intelligent Data Analysis*, Vol.6, No.5(2002), 429~250.

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).

Kubat M. and S. Matwin, "Addressing the curse of imbalanced training sets : one-sided selection", In P*roceedings of the Fourteenth International Conference on Machine Learning*, (1997), 179~186.

Laitinen, T. and M. Kankaanpaa, "Comparative analysis of failure prediction methods : the Finish case", *European Accounting Review*, Vol.8, No.1(1999), 67~92.

Laurikkala, J., "Instance-based data reduction for improved identification of difficult small classes", *Intelligent Data Analysis*, Vol.6, No.4(2002), 311~322.

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.

Meyer, P. A. and H. Pifer, "Prediction of bank failures", *The Journal of Finance*, Vol.25(1970), 853~68.

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.

Odom, M. and R. Sharda, "A neural network for bankruptcy prediction", *Proceedings of the International Joint Conference on Neural*

Networks, IEEE Press, San Diego, CA. (1990).

Ohlson, J., "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of Accounting Research*, Vol.18, No.1(1980), 109~131.

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

Pantalone, C. and M. B. Platt, "Predicting commercial bank failure since deregulation", *New England Economic Review*, (1987), 37~47.

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

Provost. F. and T. Fawcett, "Robust classification for imprecise environments", *Machine Learning*, Vol.42(2001), 203~231.

Ravi, P. and K. V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques-a review", *European Journal of Operational Research*, Vol.180(2007), 1~28.

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.

Shaw, M. and J. Gentry, "Using and expert system with inductive learning to evaluate business loans", *Financial Management*, Vol.17, No.3 (1998), 45~56.

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*, forthcoming, (2009).

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

Wu, G. and E. Chang, "Adaptive feature-space conformal transformation for imbalanced data learning", *In 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).

Yan, R., Y. Liu, and R. Hauptman, "On predicting rare classes with SVM ensembles in scene classification", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, (2003).

Zmijewski, M. E., : "Methodological issues related to the estimation of financial distress prediction models", *Journal of Accounting Research*, Vol.22, No.1(1984), 59~82.

Abstract

Ensemble Learning for Solving Data Imbalance in Bankruptcy Prediction

Myoung-Jong Kim*

In a classification problem, data imbalance occurs 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. This paper proposes a Geometric Mean-based Boosting (GM-Boost) to resolve the problem of data imbalance. Since GM-Boost introduces the notion of geometric mean, it can perform learning process considering both majority and minority sides, and reinforce the learning on misclassified data. An empirical study with bankruptcy prediction on Korea companies shows that GM-Boost has the higher classification accuracy than previous methods including Under-sampling, Over-Sampling, and AdaBoost, used in imbalanced data and robust learning performance regardless of the degree of data imbalance.

Key Words : Support Vector Machine, Under-Sampling, Over-Sampling, AdaBoost, Bankruptcy Prediction, Geometric Mean-based Boosting

* Division of Business, Dong Seo University

저 자 소 개

김명종

성균관대학교 회계학과, 동 대학원에서 경영학 석사 학위 취득 후 한국과학기술원에서 경영공학 박사학위를 취득하였다. 현재 동서대학교에 경영학부 교수로 재직하고 있으며 주요 연구관심분야는 인공지능 및 데이터마이닝, 지식경영 등이다.