지능정보연구 제14권 제3호 2008년 9월(pp.133~154)

유전자 알고리즘을 활용한 데이터 불균형 해소 기법의 조합적 활용

장영식
㈜ 휴맥스 IT팀
(ysjang@humaxdigital.com)

김종우
한양대학교 경영대학 경영학부
(kjw@hanyang.ac.kr)

허 준
SPSS Korea㈜데이터솔루션
(hoh@spss.co.kr)

데이터 마이닝 분류 문제에서 발생하는 데이터 불균형 문제는 한 범주에 속한 데이터의 수가 다른 범주에 속한 데이터의 수보다 극히 많거나 작은 경우를 말한다. 이러한 불균형 문제를 해결하기 위해 표본추출과 오분류 비용에 근거한 여러 가지 기법들이 제시되었으며, 이들 간의 성능 비교에 대한 연구들도 이루어졌다. 본 논문에서는 기존에 제시된 불균형 문제 해소기법들의 조합적 활용에 대한 타당성에 대해 살펴보고 유전자 알고리즘을 통해 그 결합 비율을 결정하여 더 좋은 성과를 낼 수 있는지에 대해 살펴보도록 한다. 소수 범주에 대한 정확성을 높이기 위해 소수 범주에 대한 F-value에 기초하여 기법들의 결합비율을 결정하고 기존 단일 기법들의 성과와 임의의 비율에 의한 격자표 형태의 결합 성과를 비교하여 결합적 활용의 타당성을 살펴본다. 이를 실증적으로 검토하기 위해서, 일반적으로 데이터 불균형 문제를 해결하기 위해 많이 사용되는 4개의 공개 데이터 집합을 이용하여 타당성 분석을 수행하였다. 분석 결과, 전체적으로 단일 기법들의 결합적 활용이 데이터 불균형 해소에 유용한 것으로 나타났다.

논문접수일 : 2008년 05월 게재확정일 : 2008년 09월 교신저자 : 김종우

1. 서론

데이터 마이닝 분류(classification) 문제를 다루는 과정 중에는 발생하는 문제 중에 하나가 데이터 불균형(imbalanced data) 문제, 즉, 목표 변수(target variable) 범주의 불균형 문제이다. 데이터 불균형 문제는 목표 변수가 이탈/정상과 같이 이항형인 경우, 두 범주에 속하는 데이터 수의 비율이 현격히 차이가 나는 경우를 의미한다(오장민, 장병탁, 2001). 한 쪽의 범주에 속한 데이터의 수가 비정상적으로 큰 경우, 기계학습 알고리즘은 전체적인 오분류를 작게 하기 위해서 다수의 범주로 패턴 분류를 많이 하게 되고, 이 경우 소수의 범주는 다수의 범주로 취급된다(Weiss and Provost, 2001). 따라서 데이터 마이닝을 이용하여, 올바른 패턴 인식 모형을 개발하기 위해서는 목표변수의 분포가 50 : 50은 아니라도, 최소한 패턴을 인식할 수 있는 수준의 비율은 유지하여야 한다. 그러나 사기 적발 (Fawcett and Provost, 1997), 불균형한 단백질 구조에서 서열 규칙을 찾아내는 사례 (Radivojac et al., 2004), 바다 표면의 위성사진을 통해서, 기름 유출이 발생되는 곳을 찾아내는 사례 (Kubat et al., 1998)나 부정한 사기 전화 통화 문제에 관한 사례(Fawcett and Provost, 1996), 문서를 해당 문자 그룹 범주에 정확하게 분류하는 사례(Lewis and Ringuette, 1994) 등과 같은 다양한

상황에서 데이터 불균형 문제는 발생하게 된다(허준, 김종우, 2007). 이와 같은 데이터 불균형 문제는 기계학습 알고리즘의 성능을 저하시키는 한 요인이기도 하다(강필성 등, 2005). 기존의 데이터 불균형 문제를 해결하기 위한 표본추출과 오분류 비율에 근거한 여러 가지 기법들이 제시되었으며, 이들간의 성능 비교 연구도 수행되었다. 하지만, 이들 기법들을 적절히 조합해서 활용할 수 있는 방안이나 기법에 대한 연구는 거의 부족한 형편이다.

일반적으로 유전자 알고리즘은 고객의 신용평가, 부도 예측 등에서 모형의 최적화를 위해 많이 사용된다. 본 논문에서는 의사결정나무 추론을 사용한 데이터마이닝 문제에서 데이터 불균형 문제를 해소하기 위한 기법들의 결합적 활용의 유용성을 검증하고, 합리적인 결합 비율을 결정하기 위한 방법으로 유전자 알고리즘을 사용 방안에 대하여 연구하도록 한다. 구체적으로 다음과 같은 연구 질문들을 고찰해 보고자 한다. 첫째, 기존에 제시되었던 불균형 데이터 해소 기법들의 조합적인 활용을 통해서 생성되는 의사결정나무의 성과를 높일 수 있는가? 둘째, 만일 불균형 해소 기법들의 조합적 활용이 유용하다면, 임의적으로 조합 비율을 결정하는 것보다 유전자 알고리즘을 활용하여 조합 비율을 결정하는 것이 더 좋은 성과를 제공하는가? 본 연구에서는 이러한 연구 질문에 답하기 위해서 일반적으로 데이터 불균형 문제에 많이 사용되는 UCI 공개 데이터, Mammography 데이터와 ELENA 프로젝트 데이터를 활용하여 분석을 수행하였다. 본 논문의 구성은 다음과 같다. 제 2장에서는 관련 연구문헌에 대한 검토로 기존의 데이터 불균형 해소 기법으로서 표본추출 기법, 오분류 비용 조정, 유전자 알고리즘에 대해서 살펴본다. 제 3장에서는 이 논문이 제시하고 있는 유전자 알고리즘을 이용한 데이터 불균형 해소 기법들의 결합적 활용 방안에 대해서 살펴본다. 제 4장에서는 실험 설계에 대하여 살펴보고, 제 5장에서 실험에 대한 결과 및 분석을 제시한다. 마지막 제 6장에서는 결론을 제시한다.

2. 관련연구

2.1 데이터불균형 해소기법

불균형 문제를 해결하기 위해서, 가장 기본적인 방법으로 표본추출을 이용한 방법이 있으며, 다른 방법으로는 오분류 비용(misclassification cost)을 조정하거나 분류의 결정 기준(decision thresholds)을 조정하는 방법이 대표적이라고 할 수 있다.

2.1.1 표본추출

표본추출(Sampling) 방법에는, 다수 범주 집단에서 임의적 표본추출을 하여, 소수 범주와 균형(balance)을 이루도록 하는 과소 표본추출(Under Sampling) 방법과 소수범주 집단을 반복적으로 복사를 하여 다수 범주 집단과 균형을 이루게 하는 과대 표본추출(Over Sampling) 방법이 있다. Japkowicz(2000)는 1차원 인공데이터를 사용하여 소수범주 집단의 반복에 의한 과대 표본추출과 다수범주 집단의 과소 표본추출을 비교하는 연구를 수행하여 이 방법이 효과적임을 보였다. 하지만 Ling and Li(1998)는 다이렉트 마케팅 문제에서 소수범주에 대한 과대 표본추출이 분류 성능을 크게 향상시키지 않음을 보였다. Chawla et al.(2002)은 단순 반복이 아닌 k-Nearest Neighbour 기법을 통해 소수 범주 주변에 인공데이터를 생성하는 방법으로 SMOTE(Synthetic Minority Over-Sampling Technique)를 제안하였으며 Chawla et al.(2003)은 부스팅 기법과 SMOTE를 결합시킨

SMOTEBoost를 제안하기도 하였다. Guo and Viktor(2004)는 앙상블 기반 학습 알고리즘과 부스팅기법을 결합한 DataBoost-IM이란 알고리즘을 사용하여 데이터 불균형 문제를 해소하고자 하였다. Kubat와 Matwin(1997)은 다수범주 데이터를 4개(safe, redundant, borderline, class-label noise)의 범주로 나누고 Tomek Link와 1-NN을 통하여 safe에 속하는 데이터만을 추출하는 일종의 과소 표본추출방법을 제안하였다. Chawla et al.(2005)는 소수범주의 정확성을 향상 시키기 위해 Buckland(1994)가 정의한 정합성 척도인 F-value를 평가척도로 하여 SMOTE와 과소 표본추출 비율을 계산하는 연구를 수행하였다. 그 외에도 오장민과 장병탁(2001)은 불균형 데이터 문제를 해결하기 위해 퍼셉트론에 기반한 부스팅 기법을 제안하였고, 강필성 등(2005)은 소수범주에 속하는 데이터가 매우 적은 경우, 데이터 불균형이 실제로 분류기의 성능에 미치는 영향을 2차원 인공 데이터를 통하여 알아보고 지금까지 제안된 방법들의 단점을 극복하고자 SVM 앙상블 기법을 적용한 과대 표본추출을 제안하였다.

2.1.2 오분류 비용을 조정하거나 분류의 결정 기준을 조정하는 방법

오분류 비용을 조정하거나 각종 가중치를 사용하여 데이터의 불균형을 해소하는 방법이다. 즉, 이는 원 데이터 구조를 그대로 유지하면서, 소수범주 오분류에 가중치를 두어, 데이터의 불균형을 해소하고자 하는 방법이다. 오분류 비용의 조정은 의사결정나무 추론 기법에서만 사용이 가능한 것이라고 할 수 있고, 그 외 로지스틱 회귀분석 등의 기법에서는 목표 변수에 가중치를 다르게 주어서, 분류가 되는 기준을 변화시켜서 불균형한 데이터의 문제를 해결할 수 있다(허준, 김종우, 2007). Domingos(1999)는 비용을 최소화하는 절차를 통해 임의의 분류기를 비용에 민감하게 만드는 방법으로 MetaCost를 제안했으며 Fan et al.(1999)은 AdaBoost 보다 누적 오분류 비용을 줄이기 위한 목적으로 AdaCost를 제안하였다. 또한 Huang et al.(2004)은 오분류 기각역의 변화에 따른 불균형 데이터의 해결을 위해, 가장 낮은 영역의 실제 정확성을 직접 통제하는 BMPM(Biased Minmax Probability Machine)을 제시하였다. 김지현과 정종빈(2004)은 오분류 비용의 비를 이용한 소수범주에 대한 복원 표본추출과 가중치 부여 방식이 단순한 표본추출에 의한 균형보다 기대비용 기준에서 더 좋은 성과를 내며 부스팅 기법 적용 시에는 주어진 자료를 그대로 이용하는 것이 좋다는 결과를 보였다.

2.2 유전자 알고리즘

유전자 알고리즘이란 자연계에 있어서 생물의 유전과 진화의 메커니즘을 공학적으로 모형화하는 것을 통해 생물이 갖는 환경에서 적응 능력을 컴퓨터를 통해서 흉내 내는 것으로, 1970년대 초기에 미시간 대학 교수인 John Holland에 의해 제안된 자연도태의 원리를 기초로 한 최적화 방법이다. 즉, 유전자 알고리즘은 자연계의 진화 현상을 기반으로 만들어진 계산 모델로서 풀고자 하는 문제에 대한 가능한 해들을 정해진 형태의 자료구조로 표현한 다음, 이들을 점차적으로 변형함으로써 점점 더 좋은 해들을 생성하게 된다. 각각의 가능한 해를 하나의 유기체 또는 개체로 보며 이들의 집합을 개체군(population)이라 한다. 하나의 개체는 보통 한 개 또는 여러 개의 염색체(Chromosome)로 구성되며, 염색체를 변형하는 연산자들을 유전 연산자라 한다. 이러한 유전자 알고리즘은 탐색 및

최적화 기계 학습의 도구로 많이 사용되고 있다(조영임, 2003). 유전자 알고리즘은 특히 다른 인공지능 기법들과 통합하여 많이 적용되어 왔는데, 이는 인공지능 모형 구축에 있어서 최적화 필요를 만족시키는데 효과적이기 때문이다. 유전자 알고리즘은 초기화(Initialization), 선택(Selection), 교배(Crossover), 그리고 돌연변이(Mutation)와 같은 절차를 통해 탐색을 수행하게 된다(이건창, 1999).

이러한 유전자 알고리즘은 데이터마이닝 분야에서 여러 가지 데이터마이닝 기법의 효과적인 결합과 입력변수 최적 조합의 결정을 위해 많이 사용되어 왔다. 홍승현 등(1999)은 기업부도 예측 모형을 중심으로 인공신경망 기법의 최적 변수 조합을 선정하기 위하여 유전자 알고리즘 기법을 이용하는 방법론을 제시하였고 이 방법론이 다른 통계기법이나 전문가에 의한 변수 선택 방법론에 비해 우수함을 보였다. 김갑식 등(2003)은 할부 금융시장에서의 고객 정보 및 할부진행과정에 대한 세부내역을 바탕으로 각기 다른 기법들로 구현된 복수개의 분류모형들을 유전자 알고리즘을 이용하여 하나의 모형으로 통합하는 방법을 통해 얻어진 신용평가모형을 제시하였다. 한인구 등(1997)은 도산 예측 문제의 다양한 분류 기법들의 통합을 지원하기 위하여 유전자 알고리즘을 사용하였는데 이 연구에서 판별분석, 인공신경망, 사례기반 추론의 선형결합 오분류 비용을 최소화시키기 위해 유전자 알고리즘을 적용하였다. 실험 결과, 유전자 알고리즘을 이용한 모형의 성과가 휴리스틱 결합 방법보다 높은 예측률을 보여 주었다. 또한, 신경식 등(1998)은 채권 등급 예측을 위해 사용한 사례기반 추론에서 가중치 값을 탐색하기 위해 유전자 알고리즘을 사용하여 예측력을 향상시킬 수 있음을 보여 주였다.

3. 유전자 알고리즘을 이용한 결합적 활용방안

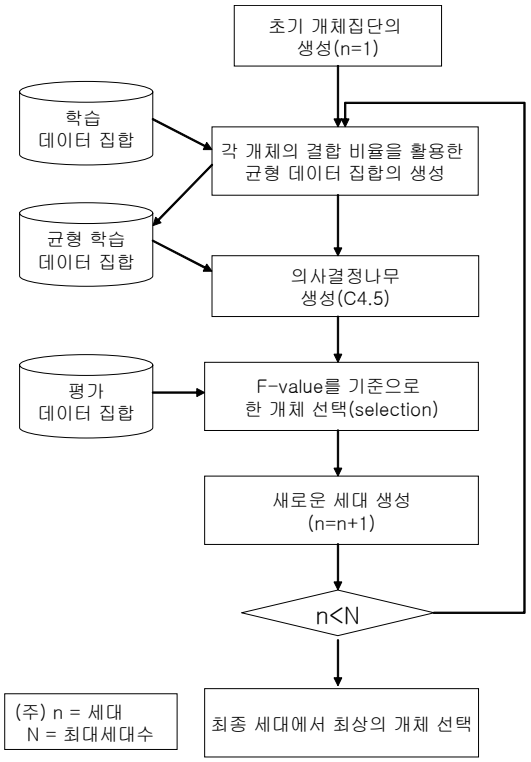
3.1 결합적 활용 방안

본 논문에서는 데이터 불균형 문제를 해소하기 위한 기법들의 결합적 활용의 유용성을 살펴보고자 한다. 여기서 결합적 활용이란, 두 가지 이상의 불균형 해소 기법들을 동시에 활용하여 데이터의 균형을 맞추는 것을 의미한다. 예를 들어, 다수 범주 집단에 대한 과소 표본추출도 하고, 동시에 소수 범주 집단에 대한 과대 표본추출을 해서 데이터의 비율의 균형을 맞추는 것을 의미한다. 하지만, 이러한 방법은 활용할 때, 다수 기법들의 활용비율을 결정하는 것은 반복적이고 주먹구구식으로 이루어지고, 최적화 조합을 얻기 위해서는 모든 가능한 조합들을 모두 고려해야 하는 데, 이것은 현실적으로 불가능하다. 따라서, 최적은 아니지만, 어느 정도 수용가능한 기법들의 조합을 결정하기 위한 체계적인 방안의 제시가 필요하다.

유전자 알고리즘은 실행가능해(feasible solution) 집합을 모두 찾아보기 어려운 경우 최선의 해(satisficing solution)을 찾는데 유용한 수단이다. 본 연구에서는 불균형 문제를 가진 데이터에서 최선의 불균형 해소 기법들의 결합 비율을 결정하기 위한 방법으로 유전자 알고리즘을 사용하고자 한다. 즉 다수의 표본추출 방식을 사용하여 새로운 학습 데이터를 생성하거나 오분류 비용을 조정하는 방법을 병행하여 활용하고 유전자 알고리즘을 통해 결합비율을 결정하는 결합적 활용 방안을 제시하고 그 유용성을 검증해 보고자 한다.

본 연구에서 불균형 데이터는 기본적으로 이항형의 형태를 가정하고 있으며, 이항형 데이터는 그 분포에 따라서 소수범주와 다수범주로 나누어

진다. 데이터 불균형을 해소하기 위한 기법으로는 표본추출 기법과 오분류를 조정하는 기법이 대표적이며, 이러한 기법들은 병행적으로 활용이 가능하다. 본 연구에서는 새로운 학습 데이터를 생성하기 위해 사용되는 방법으로 소수범주에 대한 과대표본추출, 다수 범주에 대한 과소 표본추출, 이상치 제거 후 과소 표본추출의 3가지 기법을 사용한다. 또한 표본추출에 의해 생성된 새로운 학습 집합을 학습하는 과정에서 예측력을 높이기 위해 오분류 비용의 조정을 추가하였다. 현실 세계에서 오분류 비용을 미리 알고 있는 경우는 매우 드물기 때문에 오분류 비용을 얼마로 설정할 지를 결정하는 것은 매우 힘든 일이다. 따라서 본 논문에서는 표본추출 기법들의 결합 비율 즉 표본추출 퍼센트와 오분류 비용을 결정하기 위해 유전자 알고리즘을 사용하고 예측력을 평가하기 위해서 소수범주에 대한 F-value를 적합도 함수로 사용한다. 즉, 소수범주의 과대 표본추출 비율을 R1, 다수 범주의 과소 표본추출 비율을 R2, 이상치 제거 후 과소 표본추출 비율을 R3, 오분류 비용을 R4라 하면, 결국 분류 성능을 높일 수 있는 의사결정나무를 생성할 수 있는 최선의(R1, R2, R3, R4) 조합을 구하는 것이 필요하다. 하지만, 이러한 가능한 조합들을 모두 평가하는 것은 현실적으로 불가능하므로, 유전자 알고리즘을 통해서 최선의 조합을 구하고자 한다. 각 기법의 결합 비율인 (R1, R2, R3, R4)를 유전자 또는 개체로 하고 초기 세대를 생성한다. 각 개체별로 해당 결합 비율을 활용하여 균형 학습 데이터 집합을 생성하여 C4.5 알고리즘을 활용하여 의사결정나무를 생성한다. 이 결합 비율의 성능은 평가 데이터 집합에 대한 소수범주에 대한 F-value로 판단한다. 즉, F-value를 기준으로 유전자 알고리즘의 선택 과정을 거쳐서 다시 세대를 생성하고, 이를 반복적으로 시행한다(본 연구에서는 세대수 N을 20으로 함). 최종 세대에서 소수범주에 대한 F-value가 가장 좋은 개체의 결합 비율을 최종 결합 비율로 결정한다. 본 논문에서 제시하고 있는 유전자 알고리즘을 활용한 결합 비율의 결정을 위한 흐름도는 <그림 1>과 같다.



<그림 1> 유전자 알고리즘 기반 데이터 불균형 해소 기법의 결합적 활용 흐름도

3.2 유전자 알고리즘의 매개변수 및 성과지표

유전자 알고리즘은 개체집단의 크기(Population size), 돌연변이율(Mutation rate), 교배점(Crossover point), 선택 방법(Selection method)에 따른 선택율(Selection rate) 등 여러 가지 매개변수들을 가진다. 본 논문에서 사용한 매개변수는 <표 1>과 같다.

<표 1> 유전자 알고리즘 사용을 위한 매개변수

매개변수	설정 값
유전자 수(chromosome dim)	4 (R1, R2, R3, R4)
개체군의 크기(Population size)	750
진화 횟수(max generation)	20
랜덤 선택 확률(random selection chance)	5%
교배점(crossover point)	단일점 교배
교배율(crossover rate)	0.9
돌연변이율(mutation rate)	0.05
적합도 함수(fitness function)	소수범주에 대한 F-value

유전자 또는 염색체의 수는 4개, 즉 각 기법들의 결합비율이다. 초기 모집단의 개체 크기는 750이며 이는 랜덤으로 생성된 각 기법들의 비율결합이 750개라는 것을 의미한다. 진화 횟수는 20번이며 마지막 20번째에서 최적 적합도(best Fitness)를 산출한다.

선택법에는 기본적으로 적합도 비례 룰렛휠 선택법(proportionate selection roulette selection), 엘리트 보존 선택법(Elitism), 기대치 선택법(expected-value selection), 순위 선택법(ranking selection)이 있으며 본 논문에서는 순위 선택법을 사용하였다. 순위 선택법은 미리 순위와 선택할 개체 수와의 관계를 결정해 둔다. 그리고 각 개체를 적합도 순으로 나열하고 선택할 개체를 결정해 가는 것이다(조영임, 2003). <표 1>에서 랜덤 선택 확률이 5%라는 것은 순위가 높은 것이 재생에 선택될 확률이 95%라는 것을 의미한다.

교배점은 단순 교배인 단일점 교배(One-point crossover)를 사용했으며 교배율이 0.9라는 것은 초기 개체 집단에서 10%는 그대로 남고 90%만 교배가 일어난다는 것을 의미한다. 돌연변이는 개체의 각 유전자에 대하여 일정한 돌연변이 확률을 적용하여 대립 유전자의 값을 바꾸는 것이다. 유전자 알고리즘은 개체에 근접한 새로운 개체를 생성하는 국소적인 랜덤 탐색의 일종이다. 따라서 돌연변이는 집단에서 잃어버린 유전형질을 복구하여 다양성을 유지하기 위한 수단으로 사용된다. 일반적으로 전형적인 돌연변이 확률은 0.05이하이다(조영임, 2003). 본 연구에서는 돌연변이율을 0.05로 설정하였다.

적합도 함수는 데이터 불균형 문제에서 성과 측정 방안으로 많이 사용되는 F-value(Buckland et al., 1994, Joshi et al., 2001)를 사용하며 그 결과로서 결합비율을 결정하여 성과를 평가한다. <표 2>의 교차표는 F-value 계산에 사용되며 본 논문의 경우 소수범주에 관심을 가지고 있으므로 소수범주를 양성 범주(Positive class)로 다른 범주는 음성 범주(Negative class)로 간주하였다. 본 논문에서는 소수범주 집단에 대한 예측력과 정확성을 높이기 위한 목적을 가지고 있으므로 소수범주 오분류율(FP rate), 소수범주 집단에 대한 F-value 그리고 다수 집단에 대한 F-value를 고려하여 성능을 측정하도록 한다.

<표 2> 교차표

		예측 값	
		양성	음성
실제 값	양성	TP (True Positive)	FN (False Negative)
	음성	FP (False Positive)	TN (True Negative)

F-value는 정확율(Precision)과 재현율(Recall)을 동시에 고려한 성과지표이다. 정확율은 실제 양성이나 소수범주를 예측함에 있어서 정확성을 나

타내며, 재현율은 전체 양성에 대한 정확히 예측된 소수범주 사례의 비율을 나타낸다. <표 2>를 사용하여 다음과 같이 정확율, 재현율, F-value를 계산한다.

$$precision = \frac{TP}{TP + FP} \quad (1)$$

$$recall = \frac{TP}{TP + FN} \quad (2)$$

$$F - value = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times recall + precision} \quad (3)$$

식 (3)에 β는 정확율과 재현율에 할당된 관련 중요성을 통제하며 일반적으로 1로 놓는다. 대부분의 기계학습에서 정확율 또는 재현율을 최적화하기 위한 최적 독립 매개변수 설정은 종종 대립되며 동시에 둘 다를 최적화하는 것은 어렵다. 이 두 개의 지표는 F-value에 포함되고 따라서 분류기(classifier)의 우수성은 F-value에 의해 측정될 수 있다(Ajay, 2004).

4. 실험설계

4.1 데이터 집합

본 논문에서는 기존에 제시된 불균형 문제 해소 기법들의 결합적 활용에 대한 타당성에 대해 살펴보고 유전자 알고리즘을 통해 그 결합 비율을 결정하여 더 좋은 성과를 낼 수 있는지에 대해 알아보기 위하여 일반적으로 데이터 불균형 문제를 해결하기 위해 많이 사용되는 UCI 공개 데이터, Mammography 데이터와 ELENA 프로젝트 데이터를 활용하였다. 본 연구에서 사용한 데이터 집합은 <표 3>과 같다. 자료의 *표시가 있는 것은 본래 다항형 자료이나 본 실험을 위해 인위적으로 이항형으로 만든 것들이다.

<표 3> 실험 데이터 집합

데이터 집합	크기	속성 수	다수 범주 : 소수 범주 비율
Pima	768	8	0.65 : 0.35
Phoneme*	5484	5	0.71 : 0.29
Satimage*	6435	36	0.90 : 0.10
Mammography	11183	6	0.98 : 0.02

1. Pima Indian Diabetes (Blake C. and C. Merz, 1998)는 2개의 범주와 768개의 샘플을 가진다. 이 데이터는 Arizona Phoenix 근처 사람들의 당뇨병 진단 결과를 확인하기 위해 사용되었다. 양성 범주의 수는 268(35%)이다.
2. Phoneme 데이터 집합은 ELENA 프로젝트[1] 데이터이다. 이 데이터의 목적은 콧소리와 입에서 나는 소리을 구별하기 위한 것이다. 5개의 속성을 가지고 콧소리의 사례는 3,818개, 입소리의 사례는 1,586개(29%)이다.
3. Satimage 데이터 집합(Blake C. and C. Merz, 1998)은 원래 6개의 클래스를 가진다. 본 연구에서는 이항형으로 만들기 위해 소수 범주로서 가장 적은 클래스를 사용하고 나머지 클래스들은 하나의 클래스 즉 다수 범주로 사용하였다(Provost et al., 1998). 따라서 다수 범수의 수는 5,809개 이며 소수 범주의 수는 626개(10%)이다.
4. Mammography 데이터 집합(Woods et al., 1993)은 10,923개의 음성 표본(non-calcifica-

1) ELENA 프로젝트 ftp.dice.ucl.ac.be in the directory of /pub/neural-nets/ELENA/database.

tion)과 단지 260개의 양성 표본(calcification)로 구성되어 있다. 이 경우 분류 성능의 우수성으로서 예측 정확성을 본다면 모든 표본이 non-calcification으로 분류할 때 예측 정확성은 97.68%가 될 것이다. 이 데이터 집합은 SMOTE(Chawla et al., 2002)와 SMOTEboost (Chawla et al., 2003) 연구에 활용되었다.

4.2 실험방법

본 연구에서는 (1) 데이터 불균형 해소 기법의 단독활용, (2) 데이터 불균형 해소 기법들의 결합적 사용(결합 비율을 등간격으로 부여), (3) 유전자 알고리즘을 활용한 데이터 불균형 해소기법의 결합 비율 결정의 3가지 형태의 실험이 이루어 졌다.

실험방법 (1)은 기존의 데이터 불균형 해소 기법들을 단독적으로 사용한 성과를 평가하기 위하여 <표 3>의 데이터 집합을 가지고 첫 번째, 랜덤 과대 표본추출 기법으로서 소수 범주에 대하여 120%에서 2100%까지 과대 표본추출한 경우, 두 번째, 랜덤 과소 표본추출기법으로서 다수 범주에 대하여 1%에서 100% 과소 표본추출 한 경우, 세 번째, K-Nearest Neighborhood를 이용하여 이상치를 제거한 후 다수 범주에 대하여 1%에서 100%까지 과소 표본추출 한 경우, 네 번째, 오분류 비용을 임의로 조정한 경우의 4가지 실험을 하였다. 실험방법 (1)의 오분류 비용의 최고 값은 훈련용 데이터에서 소수 범주와 다수 범주의 비율이 같아지는 비율(허명회, 이용구, 2003)의 2배까지 고려하였다. 즉 오분류 비용의 최고 값은 "(다수범주의 수/소수범주의수) × 2"와 같다. <표 4>는 데이터 불균형 해소를 위한 단독기법에 대한 설계를 요약한 것이다.

실험방법 (2)는 첫째, 각 기법들의 비율을 조합한다고 가정하고 일정한 간격의 결합을 보여주는 격자표를 작성하여 그 결과를 비교하는 실험을 하였다. 실험방법 (2)에서 격자표의 경우의 수는 기법을 사용하지 않는 경우(100 또는 0)를 포함하여 11가지이며 따라서 고려한 총 경우의 수는 11(ROS) × 11(RUS) × 11(USWO) × 11(COST) = 14641가

<표 4> 데이터 불균형 해소를 위한 단독기법에 대한 설계

방 법	설 명	샘플링 비율 (%, 방법1, 2, 3의 경우) / Maximum Cost (방법 4의 경우)	경우의 수	약 칭
1	소수 범주에 대한 과대 표본추출	100, 120~2100	101	ROS (Random Over Sampling)
2	다수 범주에 대한 과소 표본추출	0, 1~100	101	RUS (Random Under Sampling)
3	K-NN을 통해서 이상치를 제거한 다수범주에 대한 과소 표본추출	0, 1~100	101	USWO (Under Sampling without Outlier)
4	오분류 비용 조정*	Pima : 4.0	101	COST
		Phoneme : 5.0		
		Satimage : 19.0		
		Mammography : 84.0		

지이다.

실험방법 (3)은 유전자 알고리즘을 통하여 결합비율 및 오분류 비용을 결정하는 실험을 하였다. 실험방법 (3)은 개체군의 수를 750개로 하여 20세대까지 총 15000개의 경우의 수를 가진다. 그러나, 교배율이 90%이고 돌연변이율이 5% 이므로 각 세대의 5%는 다음 세대에 생존하게 된다. 따라서 실제 총 경우의 수는 실험방법 (2)의 경우의 수보다 작은 750 × 0.95 × 20세대 = 14250개의 경우의 수를 가진다. 또한 각 기법의 경우의 수는 기법을 사용하지 않는 경우(0 또는 100)를 포함하여 101로 설정하였으며 적합도 함수는 소수 범주 집단에 대한 F-value로 설정하였다.

실험방법 (1)과 실험방법 (2), 실험방법 (3)을 통해 단일 기법의 단독 사용보다 결합적 사용이 더 유용한지, 그리고 결합적 사용이 더 성과가 높다면 그 결합비율을 유전자 알고리즘을 통해 결정하는 것이 더 효과적인지 살펴보고자 3가지 실험방법의 성과를 비교 분석하였다. 이를 구현하기 위한 프로그램의 기본 API는 JAVA를 기반으로 한 WEKA[2]와 Java GALib(Genetic Algorithm Library)[3]를 활용하였다. WEKA는 'Waikato Environment for Analysis'의 약어로, 와이카토 대학에서 개발하여 공개된 자바 기반의 데이터 마이닝 도구이다.

4.2.1 실험방법(1) : 데이터 불균형 해소 기법의 단독 사용

실험방법(1)의 4가지 단독 기법의 실험 수행 절차는 구체적으로 다음과 같다.

① 실험을 위해서 <표 3>에 표시된 데이터에서 무작위로 70%를 선출해서 학습집합에 넣고, 평가집합에는 학습집합에 포함되지 않은 나머지 30%를 넣는다.

② 다음의 4가지 데이터불균형 해소 기법을 단독으로 사용하여 실험하였다.

②-a (ROS) : 학습집합의 소수 범주 데이터에 대해서 120%부터 2100%까지 100개의 등간격으로 과대 표본추출하여 생성된 새로운 학습집합 데이터를 C4.5를 통해 학습 평가하고 정확율과 재현율, F-value를 계산한다.

②-b (RUS) : 학습집합의 다수 범주 데이터에 대해서 1%부터 100%까지 과소 표본추출하여 생성된 새로운 학습집합 데이터를 C4.5를 통해 학습 평가하고 소수범주에 대한 F-value를 계산한다.

②-c (USWO) : K = 5으로 하는 KNN을 통해 학습집합의 이상치 제거를 한 후 다수 범주 데이터를 1%에서 100%까지 과소 표본추출하여 생성된 새로운 학습집합 데이터를 C4.5를 통해 학습, 평가하고 소수범주에 대한 F-value를 계산한다.

②-d (COST) : 학습집합 데이터에서 오분류비용을 최대비용까지 100개 간격으로 구분하여 변화시키면서 C4.5를 통해 학습 평가하고 소수범주에 대한 F-value를 계산한다.

③ 소수범주에 대한 F-value가 높은 상위 20개를 채택하여 소수범주 오분류율, 그리고 다수범주에 대한 F-value를 구한다.

④ 랜덤 상수를 변화시키면서 위 과정을 20회 반복하여 그 평균 값을 결과 값으로 채택한다.

2) http : //www.cs.waikato.ac.nz/ml/weka/.
3) http : //sourceforge.net.

4.2.2 실험방법(2) : 격자표를 통한 데이터 불균형 해소기법의 결합적 사용 (결합비율을 등간격으로 부여)

실험방법 (2)의 격자표에 의한 성과 검증 절차는 다음과 같다.

① 실험을 위해서 <표 3>에 표시된 데이터에서 무작위로 70%를 선출해서 학습집합에 넣고, 평가집합에는 학습집합에 포함되지 않은 나머지 30%를 넣는다.

② 4가지 기법별로 각각 11가지의 단계(기법을 적용하기 않을 경우 포함)로 등간격으로 변화시켜, 이들 조합(11^4 = 14641가지)의 격자표를 작성한 후 이러한 조합을 활용해서 생성된 새로운 학습집합 데이터를 C4.5를 통해 학습 평가하고 소수범주에 대한 F-value를 계산한다.

③ ②의 과정에서 소수범주에 대한 F-value가 높은 상위 20개를 채택하여 소수범주 오분류율 그리고 다수범주에 대한 F-value를 구한다.

④ 랜덤 상수를 변화시키면서 위 과정을 20회 반복하여 그 평균 값을 결과값으로 채택한다.

4.2.3 실험방법(3) : 유전자 알고리즘을 활용한 데이터 불균형 해소기법의 결합비율 결정

실험방법 (3)의 절차는 다음과 같다.

① 실험을 위해서 <표 3>에 표시된 데이터에서 무작위로 70%를 선출해서 학습집합에 넣고, 평가집합에는 학습집합에 포함되지 않은 나머지 30%를 넣는다.

② 4가지 기법들의 비율을 <표 4>의 범위 내에서 각각 무작위 추출한 750개를 개체군으로 하여 20세대 까지 유전자 알고리즘을 실행한다. 단, 각 기법의 비율(또는 오분류 비용)은 기법을 적용하지 않는 경우를 포함하여 101가지 중 하나로 한다. 실험에 사용된 유전자 알고리즘의 매개 변수 값은 <표 2>와 같다.

③ 20번째 세대의 최적 적합도(소수범주에 대한 F-value 기준)를 가지는 비율을 채택하여 소수범주 오분류율, 그리고 다수범주에 대한 F-value를 구한다.

④ 랜덤 상수를 변화시키면서 위 과정을 20회 반복하여 소수범주에 대한 F-value가 가장 높은 값을 결과값으로 채택한다.

5. 실험 결과와 분석

5.1 실험결과

5.1.1 Pima 데이터

5.1.1.1 실험방법(1)에 대한 결과

<표 5>은 Pima 데이터를 사용하여 제 4.2절에서 언급한 데이터 불균형 해소 기법들을 단독으로 사용하는 실험을 수행한 결과이다. 성과 평가는 소수범주 오분류율(FP), 소수범주에 대한 F-value와 다수 범주에 대한 F-value로 나타내었다. 또한 각 집합간의 차이를 분석하기 위해 일원분산분석을 이용하여 검정하였고 사후 분석으로 던칸 테스트(Duncan Test)를 수행하였다.

<표 5>을 보면 먼저 4가지 방법에 따른 3개의 지표의 차이가 있는지에 대하여 일원분산분석을 이용하여, 검정한 결과 유의수준 5%에서 유의한 것으로 나타났다. 귀무가설은 4가지 방법의 소수범주 오분류율, 소수범주에 대한 F-value, 다수 범주에 대한 F-value의 평균 값이 같다는 것이며, 검정 결과 소수범주 오분류율의 F값이 47.181, 소수범주에 대한 F-value의 F값이 53.258, 다수범주에 대한 F-value의 F값이 459.671이고 p값이 모두 0.000이어서 귀무가설은 기각되었다. 이는 4가지

<표 5> 데이터 불균형 해소 기법의 단독사용[Pima 데이터]

기 법	결합비율/ Cost조정	성과평가			F-검정(p)		
		소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value	소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value
ROS	1649	0.30677	0.61997	0.76590	47.181 (0.000)	53.258 (0.000)	459.671 (0.000)
RUS	41.35	0.22227	0.61497	0.70244			
USWO	59.35	0.21783	0.62697	0.71866			
COST	2.017	0.23161	0.64740	0.75771			
Duncan Test							
소수범주 오분류율		USWO = RUS = COST < ROS					
소수범주에 대한 F-value		RUS < ROS < USWO < COST					
다수범주에 대한 F-value		RUS < USWO< COST = ROS					

방법의 성과가 차이가 난다는 것으로 볼 수 있다.

자세히 살펴보면, 이상치를 제거한 후 과소 표본추출을 한 경우가 소수범주 오분류율이 가장 좋은 것으로 나타났고, 다음으로 과소 표본추출과 오분류 비용을 조정하는 방법이 좋은 것으로 나타났다. 소수범주에 대한 F-value 지표는 오분류 비용을 조정하는 방법이 가장 좋은 것으로 나타났으며, 다수범주에 대한 F-value 지표는 과소 표본추출 방법이 가장 좋은 것으로 나타났다. 4가지 실험 결과를 불균형해소 기법을 사용하지 않고 학습의 한 경우(Baseline)와 비교해 볼 때(<표 6> 참조) 4가지 모든 기법이 소수범주 오분류율을 낮추는 것으로 나타났다.

5.1.1.2 실험방법(2)에 대한 결과

<표 6>은 Pima 데이터를 가지고 실험방법 (2)을 실행한 결과이다. 즉, 제 4.3절에서 언급한 격자표를 통해 임의비율로 단독기법들을 결합하는 실

<표 6> 실험 결과[Pima 데이터]

실험 방법	기 법	결합비율				성과평가		
		ROS	RUS	USWO	COST	소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value
1	Baseline	-	-	-	-	0.43270	0.58640	0.80424
	ROS	1649	-	-	-	0.30677	0.61997	0.76590
	RUS	-	41.35	-	-	0.22227	0.61497	0.70244
	USWO	-	-	59.35	-	0.21783	0.62697	0.71866
	COST	-	-	-	2.017	0.23161	0.64740	0.75771
2	Matrix	960	11.5	28	2.32	0.19828	0.63700	0.72358
	증감	-	-	-	-	-0.23442	+0.0506	-0.0806
3	GAbased	684	57.5	70.8	2.116	**0.16393**	**0.71820**	**0.81280**
	증감	-	-	-	-	-0.26877	+0.1318	+0.00856

험을 수행한 결과이다. 그 결과, 격자표를 통한 단독 기법의 결합이 기법을 적용하지 않은 경우(Baseline)보다 소수범주 오분류율을 23.442% 낮추었으며 소수범주에 대한 F-value 값은 5.06% 증가하고 다수 범주에 대한 F-value 값은 8.06% 감소하였음을 알 수 있다. 이는 데이터 불균형 해소 기법의 결합이 다수 범주에 대한 F-value는 낮추지만 소수범주 오분류율을 낮추고 소수범주에 대한 F-value는 높일 수 있다고 해석할 수 있다.

5.1.1.3 실험방법 (3)에 대한 결과

<표 6>의 실험방법 (3)은 Pima 데이터를 가지고 4.3절에서 언급한 유전자 알고리즘을 통해 결합비율과 오분류 비용을 결정하고 학습하는 실험을 수행한 결과이다. 실험 결과 소수범주에 대한 F-value는 13.18% 증가하였고, 소수범주 오분류율은 26.877% 감소하였다. 또한 <표 6>의 실험방법(2) 격자표를 통한 임의비율의 결합에 대한 결과와는 달리 다수범주에 대한 F-value도 미미하지만 증가하는 것을 볼 수 있었다.

5.1.2 Phoneme 데이터

5.1.2.1 실험방법 (1)에 대한 결과

<표 7>는 Phoneme 데이터를 가지고 제 4.3절에서 언급한 실험방법 (1)의 4가지 실험을 수행한 결과이다. Phoneme 데이터의 다수 범주와 소수 범주의 분포는 71 : 29이다. <표 7>에서 F-검정 결과 소수범주 오분류율의 F값이 54.795, 소수범주에 대한 F-value의 F값이 84.867, 다수범주에 대한 F-value 값이 1233.007이고 p값이 모두 0.000이어서 귀무가설은 기각되었으며 이는 4가지 방법의 성과가 차이가 난다는 것으로 볼 수 있다. 또한 던칸 테스트 결과에서 볼 수 있듯이, Phoneme data의 경우는 오분류 비용을 조정한 경우 소수범주 오분류율이 가장 낮게 나타났으며 이 때의 오분류 비용은 1.62이다. 또한 소수 범주에 대한 F-value 값과 다수 범주에 대한 F-value값은 과대 표본추출을 한 경우 가장 높게 나타났다.

5.1.2.2 실험방법 (2)에 대한 결과

<표 8>의 실험방법 (2)는 Phoneme 데이터를

<표 7> 데이터 불균형 해소 기법의 단독사용[Phoneme 데이터]

기 법	결합비율 /Cost조정	성과평가			F-검정(p)		
		소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value	소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value
ROS	1296	0.17667	0.76836	0.89407	54.795 (0.000)	84.867 (0.000)	1233.007 (0.000)
RUS	90.5	0.20688	0.75184	0.88914			
USWO	89.95	0.18032	0.75346	0.88462			
COST	1.62	0.17254	0.76521	0.89105			
Duncan Test							
소수범주 오분류율		COST < ROS = USWO < RUS					
소수범주에 대한 F-value		RUS < USWO < COST < ROS					
다수범주에 대한 F-value		USWO < RUS < COST < ROS					

<표 8> 실험결과[Phoneme data]

실험 방법	기 법	결합비율				성과평가		
		ROS	RUS	USWO	COST	소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value
1	Baseline	-	-	-	-	0.23826	0.75912	0.89983
	ROS	1296	-	-	-	0.17667	0.76836	0.89407
	RUS	-	90.5	-	-	0.20688	0.75184	0.88914
	USWO	-	-	89.95	-	0.18032	0.75346	0.88462
	COST	-	-	-	1.62	0.17254	0.76521	0.89105
2	Matrix	420	75.5	74.5	2.42	0.17916	0.76822	0.89446
	증감	-	-	-	-	-0.0591	+0.0091	-0.00537
3	GAbased	379	80.15	85.75	1.792	**0.14671**	**0.80132**	**0.90987**
	증감	-	-	-	-	-0.09155	+0.0422	+0.01004

가지고 제 4.3절에서 언급한 임의로 격자표를 통해 단독기법을 결합한 실험을 수행한 결과이다. Phoneme 데이터의 경우 격자표를 통한 단독 기법의 결합이 Baseline보다 소수범주 오분류율이 5.91% 감소하였고 소수범주에 대한 F-value와 다수 범주에 대한 F-value 값은 각각 0.9% 증가, 0.537% 감소로 성과의 변동이 크지 않음을 알 수 있다.

5.1.2.3 실험방법(3)에 대한 결과

<표 8>의 실험방법 (3)은 Phoneme 데이터를 가지고 제 4.3절에서 언급한 유전자 알고리즘을 통해 결합비율을 결정하고 학습하는 실험을 수행한 결과이다. <표 8>에서 알 수 있듯이 격자표를 통해 임의비율로 결합하는 것보다 유전자 알고리즘을 사용하여 결합 비율을 결정한 경우 소수범주 오분류율은 약 3.2%가 더 좋게 나왔음을 알 수 있다. 또한 소수범주에 대한 F-value는 3.31% 증가하였고, 다수범주에 대한 F-value 값은 1% 증가하였다. 따라서 이 경우도 유전자 알고리즘을 사용하여 비율을 결정하는 것이 격자표를 통한 임의

<표 9> 데이터 불균형 해소 기법의 단독사용[Satimage 데이터]

기 법	결합비율 /Cost조정	성과평가			F-검정(p)		
		소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value	소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value
ROS	1640	0.39907	0.56380	0.94980	62.870 (0.000)	26.537 (0.000)	87.711 (0.000)
RUS	86.6	0.38843	0.55587	0.94700			
USWO	87.7	0.33875	0.57008	0.94554			
COST	12.394	0.33576	0.56776	0.94678			
Duncan Test							
소수범주 오분류율		COST = USWO < RUS = ROS					
소수범주에 대한 F-value		RUS < ROS < COST < USWO					
다수범주에 대한 F-value		COST = USWO < RUS < ROS					

비율의 결합보다 성과가 좋다는 것을 알 수 있었다.

5.1.3 Satimage 데이터

5.1.3.1 실험방법(1)에 대한 결과

<표 9>은 Satimage 데이터를 가지고 제 4.3절에서 언급한 4가지 실험을 수행한 결과이다. Satimage 데이터의 경우 다수범주와 소수범주의 분포는 90 : 10로서 데이터 불균형 정도가 앞의 두 데이터 집합보다는 심한 경우이다. <표 9>에서 볼 수 있듯이, 측정결과 소수범주 오분류율의 F값이 62.870, 소수범주에 대한 F-value의 F값이 26.537, 다수범주에 대한 F-value의 F값이 87.711이고 p값이 모두 0.000이어서 귀무가설은 기각되었다. 이는 4가지 방법의 성과가 차이가 난다는 것으로 볼 수 있다.

5.1.3.2 실험방법 (2)에 대한 결과

<표 10>의 실험방법 (2)는 Satimage 데이터를 가지고 격자표를 통해 단독기법을 결합한 실험을 수행한 결과이다. <표 10>을 보면 Baseline에 비해 소수범주 오분류율과 소수범주에 대한 F-value의 성과는 좋아진 반면 다수범주에 대한 F-value는 감소했음을 알 수 있다.

5.1.3.3 실험방법 (3)에 대한 결과

<표 10>의 실험방법 (3)은 Satimage 데이터를 가지고 유전자 알고리즘을 통해 결합비율을 결정하고 학습하는 실험을 수행한 결과이다. <표 10>에서 알 수 있듯이, 유전자 알고리즘을 사용하여 결합 비율을 결정한 경우 격자표를 통한 임의비율의 결합보다 소수범주 오분류율은 4.94%, 소수범주에 대한 F-value는 6.739%의 성과가 향상되었음을 알 수 있다. 또한 다수 범주에 대한 F-value도 감소되지 않았음을 알 수 있다.

5.1.4 Mammography 데이터

5.1.4.1 실험방법 (1)에 대한 결과

<표 11>은 Mammography 데이터를 가지고 제 4.3절에서 언급한 4가지 단독기법의 사용의 실험결과이다. Mammography 데이터의 경우 다수범주와 소수범주의 분포는 98 : 2로서 데이터 불균형 정도가 아주 심한 경우이며 본 논문에서 가장 관

<표 10> 실험결과[Satimage 데이터]

실험 방법	기 법	결합비율				성과평가		
		ROS	RUS	USWO	COST	소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value
1	Baseline	-	-	-	-	0.46798	0.55773	0.95495
	ROS	1640	-	-	-	0.39907	0.56380	0.94980
	RUS	-	86.6	-	-	0.38843	0.55587	0.94700
	USWO	-	-	87.7	-	0.33875	0.57008	0.94554
	COST	-	-	-	12.394	0.33576	0.56776	0.94678
2	Matrix	1240	60.5	75.5	4.69	0.32865	0.57584	0.92797
	증감	-	-	-	-	-0.13933	+0.01811	-0.02698
3	GAbased	758	79.5	79.4	5.005	**0.27926**	**0.64323**	**0.95650**
	증감	-	-	-	-	-0.18873	+0.0855	+0.00155

<표 11> 데이터 불균형 해소 기법의 단독사용[Mammography 데이터]

기 법	결합비율 /Cost조정	성과평가			F-검정(p)		
		소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value	소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value
ROS	359	0.41154	0.60655	0.99093	61.012 (0.000)	33.748 (0.000)	61.778 (0.000)
RUS	84.9	0.43297	0.60537	0.99119			
USWO	86.3	0.40000	0.60954	0.99086			
COST	12.454	0.37552	0.58546	0.98944			
Duncan Test							
소수범주 오분류율		COST < USWO < ROS < RUS					
소수범주에 대한 F-value		COST < RUS = ROS = USWO					
다수범주에 대한 F-value		COST < USWO = RUS < ROS					

심이 되는 경우이다. <표 11>에서 알 수 있듯이, 일원분산분석을 이용하여, 검정한 결과 유의수준 5%에서 유의한 것으로 나타났다. 이는 4가지 방법의 성과가 차이가 난다는 것으로 볼 수 있다. 자세히 살펴보면, 오분류 비용을 조정한 경우가 소수범주 오분류율이 가장 좋은 것으로 나타났다. 하지만 오분류 비용을 조정한 경우 다른 두 가지 지표에서는 성과가 가장 좋지 않음을 알 수 있다. 소수범주에 대한 F-value는 이상치 제거 후 과소 표본추출, 랜덤 과대 표본추출, 랜덤 과소 표본추출 방법이 비슷한 성과를 나타냄을 알 수 있다.

5.1.4.2 실험방법 (2)에 대한 결과

<표 12>의 실험방법 (2)는 Mammography 데이터를 가지고 격자표를 통해 임의비율로 단독기법을 결합한 실험을 수행한 결과이다. 실험 결과 Baseline보다 소수범주 오분류율이 5.35% 감소하는 것 외에 다른 지표에서는 거의 비슷한 성과를

<표 12> 실험결과[Mammography data]

실험 방법	기 법	결합비율				성과평가		
		ROS	RUS	USWO	COST	소수범주 오분류율	소수범주에 대한 F-value	다수범주에 대한 F-value
1	Baseline	-	-	-	-	0.47779	0.60772	0.99205
	ROS	359	-	-	-	0.41154	0.60655	0.99093
	RUS	-	84.8	-	-	0.43297	0.60537	0.99119
	USWO	-	-	86.3	-	0.40000	0.60954	0.99086
	COST	-	-	-	12.454	0.37552	0.58546	0.98944
2	Matrix	150	40	66.5	1	0.42429	0.60864	0.99122
	증감	-	-	-	-	-0.0535	+0.00092	-0.00083
3	GAbased	434	80.15	86.5	4.9425	**0.30808**	**0.70404**	**0.99306**
	증감	-	-	-	-	-0.16971	+0.09632	+0.001

보이는 것으로 나타났다.

5.1.4.3 실험방법(3)에 대한 결과

<표 12>의 실험방법(3)은 Mammography 데이터를 가지고 제 4.3절에서 언급한 유전자 알고리즘을 통해 결합비율을 결정하고 학습하는 실험을 수행한 결과이다. <표 12>의 결과를 비교해 보면 소수범주 오분류율은 16.97% 감소하였고 소수범주에 대한 F-value는 9.62% 증가하였다. 따라서 다수범주에 대한 F-value는 감소하지 않으면서 소수범주에 대한 성과는 향상되었음을 알 수 있다.

5.2 실험분석

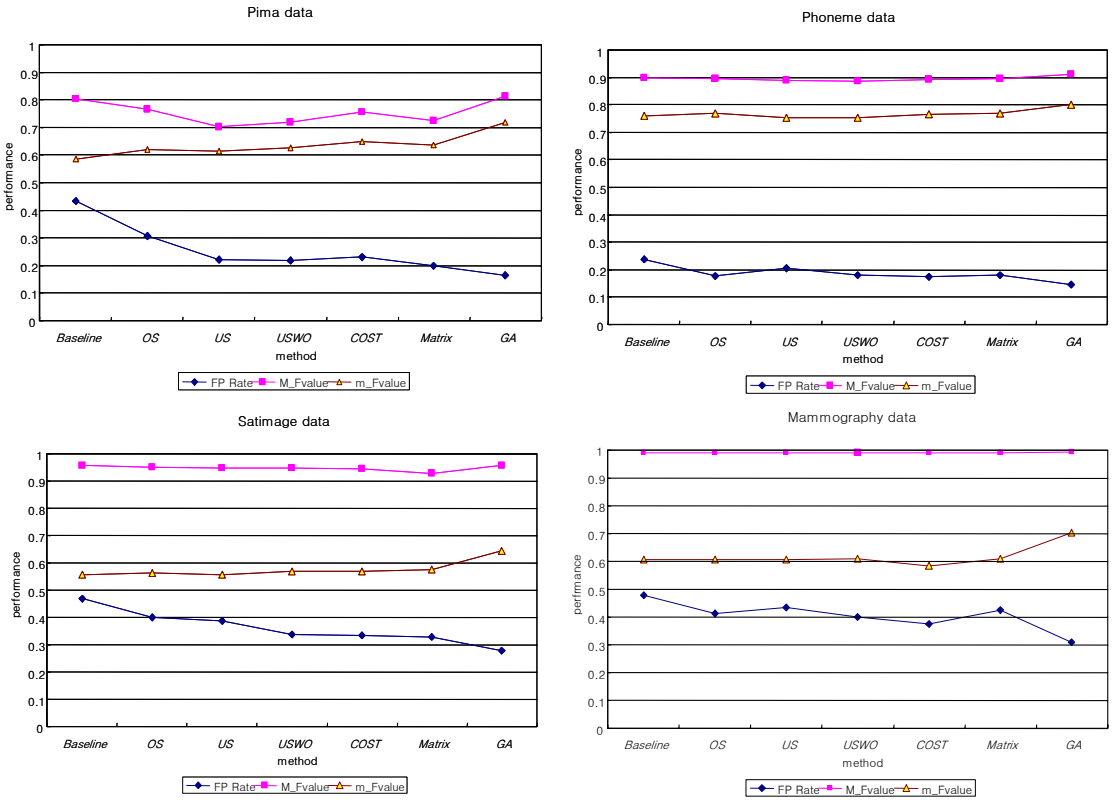
4개의 데이터 집합을 활용한 실험을 통해서 얻어진, 본 연구에서 검토하고자 한 연구 질문에 대한 검증 결과는 다음과 같다.

첫째, 기존에 제시되었던 불균형 데이터 해소 기법들의 조합적인 활용을 통해서 생성되는 의사결정나무의 성과를 높일 수 있는 것으로 나타났다. 모든 데이터 집합에서 단일 기법을 사용하는 경우보다 격자표를 사용하는 경우에 소수 범주에 대한 F-value가 높게 나타났으며 소수범주 오분류율도 낮게 나타났다.

둘째, 격자표를 통해 임의적으로 조합 비율을 결정하는 것보다 유전자 알고리즘을 활용하여 조합 비율을 결정하는 것이 더 좋은 성과를 제공하는 것으로 나타났다. 본 논문에서는 모든 데이터 집합에서 격자표에 의한 임의 비율보다는 유전자 알고리즘을 이용한 조합 비율에서 소수범주와 다수 범주에 대한 F-value, 그리고 소수범주 오분류율이 높게 나타남을 보였다.

<그림 2>는 각 데이터에 대한 결과를 종합한 것이다. <그림 2>에서 FP Rate은 소수범주 오분류율이며, M_Fvalue는 다수범주에 대한 F-value, 그리고 m_Fvalue는 본 논문에서 적합도함수로 사용한 소수범주에 대한 F-value를 나타낸다. <그림 2>의 Pima Data에 대한 실험 결과를 보면, 소수범주 오분류율을 고려할 때, 실험방법 (1)보다 실험방법 (2)가, 실험방법 (2)보다는 실험방법 (3)의 성과가 더 높다는 것을 알 수 있다. 또한 소수범주에 대한 F-value의 경우, 실험방법 (1)과 실험방법 (2)의 성과가 크게 차이가 없으며 실험방법 (3)의 성과가 가장 높다는 것을 알 수 있다. 일반적인 데이터 불균형 해소를 위한 기법들의 경우 소수 범주에 대한 F-value가 높아지면 다수 범주에 대한 F-value가 감소하는데 반해 Pima 데이터에서는 유전자 알고리즘을 통해 결합비율을 결정함으로써 다수 범주에 대한 F-value를 낮추지 않고도 단독 기법들의 결합 성과가 높게 나타났음을 볼 수 있다.

<그림 2>의 Phoneme 데이터에 대한 각 실험 결과를 보면, 성과평가 방법 3가지 모두를 고려할 때. 실험방법 (1)보다는 실험방법 (2)가, 실험방법 (2)보다는 실험방법 (3)의 성과가 더 높은 것을 볼 수 있다. 실험방법 (1)에서 오분류 비용 조정과 과대 표본추출 기법을 사용하는 것이 가장 예측력이 높은 것으로 나타났으며 전체적으로 볼 때 소수범주 오분류율은 Baseline보다 좋게 나타났다. <그림 2>의 Satimage 데이터에 대한 각 실험 결과를 보면, 소수범주 오분류율과 소수 범주에 대한 F-value를 고려할 때, 실험방법 (1)보다는 실험방법 (2)가, 실험방법 (2)보다는 실험방법 (3)의 결과값이 더 높다는 것을 알 수 있다. 또한 실험방법(2)의 격자표를 통해 임의비율로 기법들을 결합하는 것은 성과가 거의 차이가 없거나 낮아지는 경우가 있음을 확인할 수 있다. <그림 2>의 Mammography 데이터에 대한 각 실험 결과를 보면, 소수범주에 대한 F-value를 고려할 때, 실험방법 (1)보다는 실험방



<그림 2> 실험 종합

법 (2)가, 실험방법 (2)보다는 실험 (3)의 결과 값이 더 높다는 것을 알 수 있다. 하지만 FP Rate의 경우, 실험방법 (3)의 결과가 가장 좋으나, 실험방법 (2)의 결과가 실험방법(1)에 비해서 못하는 경우가 발생하였다. 따라서 <그림 2>의 실험 결과를 종합해 보면, 격자표를 통해 임의 비율로 데이터 불균형 해소 기법들을 결합하는 것은 한계가 있으며 유전자 알고리즘을 이용하여 기법들의 결합비율을 결정하는 것이 성과가 더 높다는 것을 확인할 수 있다.

6. 결론

본 연구에서는 기존에 제시되었던 불균형 데이터 해소 기법들, 즉 다수의 샘플링 방식을 사용하여 새로운 학습 데이터를 생성하거나 오분류 비용을 조정하는 방법을 병행하여 활용하고 유전자 알고리즘을 통해 결합비율을 결정하는 결합적 활용 방안을 제시하였다. 기존 단일 기법들의 성과와 임의의 비율에 의한 결합 성과를 비교하기 위해 격자표를 작성하여 비교하였다. 또한 소수 범주에 대한 정확성을 높이기 위해 소수 범주에 대한 F-value를 유전자 알고리즘의 적합도함수(Fitness function)로 하여 불균형 해소 기법들의 결합비율을 유전자 알고리즘을 이용하여 결정하고 그 성과를 측정하였다. 4개의 공개된 데이터 집합을 활용하여 분석한 결과, 전체적으로 기존 단일 기법들의

성과보다는 격자표에 의한 결합에 따른 성과가 더 높게 나타났으며 활용할만한 성과를 보였다. 또한 유전자 알고리즘을 통해 데이터 불균형 해소 기법들의 결합 비율 결정 방법이 격자표에 의한 결합보다 성과가 더 좋다는 것을 확인하였다. 따라서 데이터 불균형 해소를 위해 단일 기법의 사용보다는 결합적으로 활용하는 것이 유용하며 단일 기법들의 결합 비율을 유전자 알고리즘을 통해 결정하는 것이 성과를 더 높일 수 있음을 확인하였다.

본 연구의 한계와 향후 연구 과제는 다음과 같다. 본 연구에서는 일반적인 단순 샘플링 기법을 사용하였지만 SMOTE 등과 같은 데이터 불균형을 해결하기 위한 다양한 샘플링 방식이 연구되었다. 따라서 기존에 연구된 다양한 샘플링 방식을 구현하여 그 결과를 비교하는 것이 필요하다. 또한 본 연구에서는 일반적으로 데이터 불균형 문제를 해결하기 위해 사용되는 4개의 공개 데이터 집합을 활용하였다. 하지만, 좀 더 다양한 분포를 가지는 데이터를 통한 추가적인 실험이 요구된다. 또한 유전자 알고리즘의 성과가 매개변수의 결정에 따라서 영향을 받으므로, 매개변수를 다양하게 변화시켜서 성과를 비교 평가하는 것도 필요하다.

참고문헌

강필성, 이형주, 조성준, “데이터 불균형 문제에서의 SVM 앙상블 기법의 적용”, *한국정보과학회 가을 학술발표논문집*, (2005), 706~708.

김지현, 정종빈, “계급 불균형 자료의 분류 : 훈련표본 구성방법에 따른 효과”, *응용통계연구*, 17권 3호(2004), 445~457.

김갑식, 이동만, 황하진, “유전자 알고리즘을 이용한 할부금융회사의 고객 신용평가 데이터마이닝 모형구축”, *Journal of Business Research*, 18권 4호(2003), 249~272.

신경식, 한인구, “A Hybrid Approach Using Case-based Reasoning and Genetic Algorithms for Corporate Bond Rating”, 한국지능정보시스템학회, *한국지능정보시스템학회 학술대회*, (1998), 106~109.

오장민, 장병탁, “불균형 데이터의 효과적 학습을 위한 커널 퍼셉트론 부스팅 기법”, *한국정보과학회 춘계학술발표논문집(B)*, (2001), 304~306.

이건창, “사례기반추론과 유전자 알고리즘을 결합한 지식경영 방법론에 관한 연구 : 신용평가를 중심으로”, *정보기술응용연구 창간호*, 1999.

조영임, 인공지능시스템, 홍릉과학출판사, 2003.

한인구, 조홍규, 신경식, “The Hybrid System for Credit Rating”, *한국경영과학회지*, 22권 3호(1997), 163~173.

허명회, 이용구, 데이터 마이닝 모델링과 사례, SPSS 아카데미, 2003.

허준, 김종우, “불균형 데이터 집합에서의 의사결정나무 추론 : 종합병원의 건강 보험료 청구 심사 사례”, *Information Systems Review*, 9권 1호(2007).

홍승현, 신경식, “유전자 알고리즘을 활용한 인공신경망 모형 최적 입력변수의 선정”, *한국지능정보시스템학회 추계학술대회 논문집*, (1999), 227~249.

Ajay D. Joshi, Applying the Wrapper Approach for Discovery of Under-Sampling and Over-Sampling Percentages on Skewed Datasets, Partial Fulfillment of the Requirements for the Degree of Master, Department of Computer Science and Engineering, University of South Florida(2004).

Blake, C. and C. Merz, UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, Department of Information and Computer Science, Univer-

sity of California, Irvine(1998).

Buckland, M. and F. Gey, "The Relationship between Recall and Precision", *Journal of the American Society for Information Science*, Vol.45, No.1(1994), 12~19.

Chawla. N, A. Lazarevic, L. Hall, K. Bowyer, "SMOTEBoost : Improving Prediction of Minority Class in Boosting", 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia(2003), 107~119.

Chawla, N., V. Kevin. W. Boywer, Lawrence, O. Hall, and W. Philip Kegelmeyer, "SMOTE : Synthetic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16(2002), 231~357.

Chawla Nitesh V., Lawrence O. Hall, Ajay Joshi, "Wrapper-based Computation and Evaluation of Sampling Methods for Imbalanced Datasets", Conference on Knowledge Discovery in Data archive Proceedings of the 1st international workshop on Utility-based data mining 2005, Chicago, Illinois(2005), 24~33.

Domingos. P., "MetaCost : A General Method for Making Classifiers Cost-Sensitive", Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, ACM (1999), 155~164,

Fan W., S. Stolfo, J. Zhang, and P. Chan, "AdaCost : Misclassification Cost-Sensitive Boosting", Proceedings of the 16th International Conference on Machine Learning(1999), 97~105.

Fawcett, T. and F. Provost, "Combining Data Mining and Machine Learning for Effective User Profile", Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR. AAAI (1996), 8~13.

Fawcett, T. and F. Provost, "Adaptive Fraud Detection", *Data Mining and Knowledge Discovery*, Vol.1(1997), 291~316.

Guo, H., and H. L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation : The DataBoost-IM Approach", *SIGKDD Explorations*, Vol.6, No.1(2004), 30~39.

Huang, Kaizhu, Haiqin Yang, Irwin King, and Michael R. Lyu, "Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine.", Proceedings of the '04' IEEE Computer society conference on computer vision and pattern recognition (CVPR '04) (2004), 558~563.

Japkowicz, Nathalie., "The Class Imbalance Problem : Significance and Strategies", In Proceedings of the 2000 International Conference on Artificial Intelligence(2000).

Joshi, M., V. Kumar, R. Agrawal. "Evaluating Boosting Algorithms to Classify Rare Classes : Comparison and Improvements", First IEEE International Conference on Data Mining, San Jose, CA(2001).

Kubat M., Robert C. Holte and Stan Matwin, "Machine Learning for the Detection of Oil Spills in Satellite Radar Images", *Machine Learning*, Vol.30(1998), 195~215.

Kubat M., and S. Matwin. "Addressing the Curse of Imbalanced Training Sets : One Sided Selection". In Proceedings of the Fourth International Conference on Machine Learning(1997), 179~186.

Leung, Y., G. Li , and Z. B Xu, "A Genetic Algorithm for the Multiple Destination Routing Problems", IEEE Transactions on Evolutionary Computation,Vol.2, No.4(1998), 150~161.

Lewis, D. and Marc Ringuette, "A Comparison of

Two Learning Algorithms for Text Categorization.", Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval(1994), 81~93.

Ling D. X. and C. Li. "Data Mining for Direct Marketing : Problems and Solutions", Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) (1998).

Mostafa, M. E. and S. M. A. Eid, "A Genetic Algorithm for Joint Optimization of Capacity and Flow Assignment in Packet Switched Networks", Seventeenth National Radio Science Conference(2000) C5-2~5-6.

Pan, H. and I. Y. Wang, "The Bandwidth Allocation of ATM through Genetic Algorithm", Proceedings of IEEE GLOBECOM'91(1991), 125~129.

Provost, F., Fawcett, T., and Kohavi, R. ,"The Case Against Accuracy Estimation for Comparing Induction Algorithms", Proceedings of the Fifteenth International Conference on Machine Learning, Madison, WI. Morgan Kauffmann(1998) 445~453.

Radivojac, P., Nitesh V. Chawla, A. Keith Dunker, and Zoran Obradovic, "Classification and Knowledge Discovery in Protein Databases", *Journal of Biomedical Informatics*, Vol.37 (2004), 224~239.

Weiss, G. M., and F. Provost, "The Effect of Class Distribution on Classifier Learning", Technical Report, Department of Computer Science, Rutgers University(2001).

Woods, K., C. Doss, K. Bowyer, J. Solka, C. Priebe, and P. Kegelmeyer, "Comparative Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 7, No.6(1993), 1417~1436.

Xiawei, Z., C. Changjia, and Z. Gang, "A Genetic Algorithm for the Multicasting Routing Problems", International Conference Communication Technology Proceeding, WCC-ICCT 2000(2000) 1248~1253.

Abstract

Combined Application of Data Imbalance Reduction Techniques Using Genetic Algorithm

Young Sik Jang* · Jong Woo Kim** · Joon Hur***

The data imbalance problem which can be uncounted in data mining classification problems typically means that there are more or less instances in a class than those in other classes. In order to solve the data imbalance problem, there has been proposed a number of techniques based on re-sampling with replacement, adjusting decision thresholds, and adjusting the cost of the different classes. In this paper, we study the feasibility of the combination usage of the techniques previously proposed to deal with the data imbalance problem, and suggest a combination method using genetic algorithm to find the optimal combination ratio of the techniques. To improve the prediction accuracy of a minority class, we determine the combination ratio based on the F-value of the minority class as the fitness function of genetic algorithm. To compare the performance with those of single techniques and the matrix-style combination of random percentage, we performed experiments using four public datasets which has been generally used to compare the performance of methods for the data imbalance problem. From the results of experiments, we can find the usefulness of the proposed method.

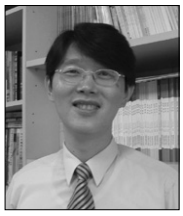
Key Words : Data Imbalance, Genetic Algorithm, Decision Tree Induction; Sampling, Misclassification Cost

* IT team, HUMAX Co., Ltd.
** School of Business, Hanyang University
*** SPSS Korea Data Solution Inc.

저 자 소 개

장영식

한양대학교 경영학부에서 경영학사(2004), 한양대학교 대학원 경영학과에서 경영학 석사(2007)를 취득하였다. 현재 ㈜ 휴맥스 IT팀에서 SAP FI(Financial Accounting) Consultant 로 재직 중이며 SAP BI(Business Intelligence) 업무를 겸하고 있다. 주요 관심분야는 데이터마이닝, 전사 자원 관리, SCM, 지능정보시스템, 의사결정지원시스템, 비즈니스 프로세스모델링 등이다.



김종우

서울대학교 수학과에서 이학사(1989), 한국과학기술원 경영과학과에서 공학석사(1991), 한국과학기술원 산업경영학과에서 공학박사(1995)를 취득하였다. 현재 한양대학교 경영대학 경영학부 교수로 재직 중이다. 충남대학교 통계학과 부교수, University of Illinois at Urbana-Champaign 방문 연구원 등을 역임하였다. 주요 관심분야는 상품/컨텐츠 추천 시스템, 데이터마이닝, 지능정보시스템, 의사결정지원시스템, 비즈니스 프로세스모델링 및 통합, 데이터 품질 등이다.

허 준

중앙대학교 정경대학 응용통계학과에서 경제학사(1998), 중앙대학교 대학원 통계학과에서 경제학 석사(2000), 그리고 한양대학교 대학원 경영학과에서 경영학 박사(2008) 학위를 취득하였다. 현대정보기술(HIT)과 현대/기아 자동차 그룹 계열사인 오토에버닷컴에서 근무를 하였고, 현재 SPSS Korea 컨설팅팀 수석연구원으로 재직 중이다. 주요 관심분야로는 데이터 마이닝과, 데이터 베이스, Business Intelligence, CRM 관련분야 및 수요예측 모델링 분야이다.