응용통계연구 제 19권 2호, 2006년, pp. 291-303

의사결정 규칙을 이용한 데이터 통합에 관한 연구*

김순영[1] 정성석[2]

요 약

대용량의 데이터로부터 의미있는 지식을 찾는 과정에서 데이터의 질은 무엇보다도 중요하다. 본 연구에서는 데이터의 충실도를 높이기 위한 방법으로 여러 경로로부터 수집된 데이터의 정보를 활용하기 위해 데이터 마이닝 알고리즘인 의사결정 규칙을 이용한 데이터 통합 기법을 제안하고, 실제 데이터를 이용하여 모의실험을 통해 제안된 알고리즘의 효율성을 비교하였다. 실험결과 제안된 알고리즘이 데이터 통합의 성능을 향상시킴을 알 수 있었다.

주요용어: 데이터통합, 통계적 매칭, 데이터보강, 데이터마이닝, k-최근접이웃, 의사결정나무, 수령자 파일, 제공자 파일.

1. 서론

일반적으로 데이터 마이닝(data mining)은 대용량의 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화함으로써 유용한 지식을 추출하는 일련의 과정들을 말하며 이는 사회 전반에 걸쳐 많은 분야에서 관심을 가지고 활발한 연구와 응용이 진행되고 있다. 이러한 과정에서 무엇보다도 가장 중요한 요소 중 하나는 데이터의 질이다. 데이터의 충실도가 만족되지 않는 상황에서의 데이터 마이닝이란 별 의미가 없으며 기대하는 결과도 얻을 수 없다. 여기서 데이터의 충실도란 데이터의 정확도, 데이터의 양(레코드 수), 데이터의 깊이(항목의 수)에 의해 평가된다. 그러나 현실에서 우리가 필요로 하는 변수를 모두 포함하는 질 좋은 데이터를 얻는 것은 어렵고, 많은 시간과 노력이 소요하게 된다. 이런 데이터 수집에 대한 어려움은 데이터 보강(data enrichment)에 의해서 해결할 수 있으며, 데이터 보강을 위해 데이터 통합(data fusion)기법을 사용할 수 있다.

본 연구에서는 데이터 보강을 위하여 여러 경로로부터 수집된 데이터를 결합하여 데이터의 깊이를 늘리기 위한 방법인 데이터 통합(data fusion) 기법에 관하여 논의하고자 한다.

* 본 연구는 산업자원부 주관 핵심기술개발연구과제로 수행되었음(과제번호: 105077001).

1) (561-756) 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보과학과, 박사과정
E-mail: rabbit@chonbuk.ac.kr

2) (교신저자) (561-756) 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교
수학통계정보과학부(응용통계연구소), 교수
E-mail: sschung@chonbuk.ac.kr

Saporta(2002)는 데이터 통합을 보유하고 있는 데이터 파일에 필요한 변수가 없거나, 결측치가 존재할 경우 다른 원천으로부터 모아지는 데이터와 정보(information)를 통합시키는 것이라고 정의했다.

데이터 통합 기법에 관한 기존 연구들을 살펴보면 근본적으로 거리(distance)와 같은 유사성(similarity) 측도를 이용하여 가장 유사한 개체(Nearest Neighbor)를 찾음으로써 이루어진다. 그리고 더 나아가서는 회귀분석(regression)을 적용한 데이터 통합 기법이 제안되었고, 최근에는 정성석 등(2004)이 회귀분석 기법에 k-최근접이웃방법(k-NN)기법을 적용하여 상대적으로 유사한 개체에 대한 정보의 손실을 줄이는 방법을 제안하였으며, 김순영 등(2005)은 군집화(clustering) 기법을 적용하여 데이터 통합의 효율성을 높이는 방법을 제안하였다.

본 연구에서는 기존의 데이터 통합에 사용되었던 통계적 기법 외에 데이터 마이닝 알고리즘으로 사용되고 있는 기계학습(machine learning)기법 중 의사결정 규칙(decision rule)을 이용한 데이터 통합 알고리즘을 제안하고자 한다.

기존의 대부분의 데이터 통합에 관한 연구는 양적 자료를 대상으로 통합시키는 기법에 한정되어 있었지만 제안하고자 하는 의사결정 규칙을 이용한 통합 알고리즘은 양적 자료뿐만 아니라 질적 자료도 적용 가능한 기법이다. 즉, 데이터의 측도에 관계없이 데이터 통합을 일괄적으로 적용시킬 수 있을 뿐 아니라, 모집단의 분포에 대한 가정이 거의 필요 없는 비모수적인 방법이 적용되며 기존의 통합 알고리즘에 비해 데이터 통합의 정확도를 높이고자 하였다. 실제 여러 데이터에 기존의 통합 알고리즘과 제안하는 통합 알고리즘을 적용시켜 통합을 수행한 결과 본 연구에서 제안하는 의사결정 규칙을 이용한 통합 알고리즘이 보다 정확한 작업을 수행함을 알 수 있었다.

데이터 보강을 위한 데이터 통합의 개념과 데이터 통합을 위한 데이터 구조에 관한 모든 설명은 정성석 등(2004)에 기술되어 있다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 기존의 데이터 통합 기법과 본 연구에서 제안하는 의사결정 규칙을 이용한 데이터 통합기법을 기술하였으며, 3절에서는 제안한 방법을 실제 데이터에 적용하여 그 성능을 기존의 방법과 비교하였다. 마지막으로 4절에서는 결론과 향후 연구방향에 대해 논의하였다.

2. 데이터 통합 기법

일반적인 데이터 통합원리는 두 파일간의 개체 간 유사성을 기준으로 이루어진다. 즉, 수령자 파일의 한 개체와 제공자 파일의 모든 개체들 간의 거리를 계산한 후, 그중 가장 가까운 거리를 갖는 제공자 파일의 개체를 선택하여 수령자 파일의 해당 개체에 추가시킨다. 이와 같은 과정을 수령자 파일의 모든 개체에 관하여 수행하여 통합된 데이터 파일을 형성한다. 이때 유사성의 측도는 유클리디안 거리(Euclidean distance)를 흔히 사용한다.

Putten et al.(2002)에 의해 제시된 데이터 통합 알고리즘이 유용한 결과를 도출하기 위한 여러 제약 조건 중 가장 중요하게 부각되는 조건은 공통변수 X가 주어졌을 때, 수령자 파일의 유일변수인 Y와 제공자 파일의 유일변수 Z사이에 조건부 독립관계가 성립되어야 한다

는 조건부 독립성이다. 조건부독립성은 통계적 매칭에서 유용한 가정으로 Rässler(2002)는 이를 판단하기 위한 방법으로 회귀분석접근법(regression approach)을 제시하였다.

본 절에서는 기존 연구의 통계적 매칭 알고리즘인 k-최근접이웃방법, 회귀분석방법, 정성석 등(2004)이 제안한 수정된 데이터 통합 기법과 김순영 등(2005)이 제안한 군집화를 이용한 데이터 통합기법을 살펴보고, 본 연구에서 제안하는 의사결정 규칙을 이용한 데이터 통합 기법에 대해 살펴보겠다.

2.1. k-최근접이웃기법

최근접이웃방법은 공통변수 X를 이용하여 수령자 파일의 각 개체에 대해 제공자 파일의 모든 개체와의 거리를 계산하여 이중 가장 가까운 제공자 파일의 하나의 개체를 이용하여 통합을 수행하는 방법이고, Putten et al.(2002)에 의해 제시된 k-최근접이웃기법은 최근접이웃방법과 마찬가지로 수령자 파일의 각 개체에 대해 제공자 파일의 모든 개체와 거리를 계산한 후, 상대적으로 유사한 제공자 파일의 k개의 개체를 선택하고 선택된 k개 개체에 해당하는 제공자 파일의 유일변수 Z를 이용하여 수령자 파일의 각 개체에 통합 변수를 추가시킨다. 이때 유일변수가 연속형인 경우 k개 Z값의 평균을, 범주형이면 k개 Z값의 최빈값를 이용한다.

2.2. 회귀분석 알고리즘

Ingram et al.(2000)에 의해 제시된 회귀분석을 이용하는 기법은 제공자 파일에서 회귀모형을 추정한 후, 추정된 회귀모형을 이용하여 수령자 파일과 제공자 파일에서 예측치를 구하고, 두 파일의 개체간 예측치 사이의 거리가 가장 가까운 개체를 이용하여 통합에 사용하는 방법이다. 이 방법을 자세히 살펴보면 다음과 같다.

제공자 파일을 대상으로 공통변수 X를 설명변수로 하고 유일변수 Z를 반응변수로 하는 회귀모형을 추정한 후, 제공자 파일에서 추정된 회귀모형을 수령자 파일과 제공자 파일에 적용하여 각 파일에서 유일변수 Z의 예측값 $\hat{Z}$를 구한다. 이 예측값 $\hat{Z}$을 이용하여 수령자 파일의 각 개체에 대해 제공자 파일의 모든 개체사이의 거리를 계산하여 거리가 가장 가까운 제공자 파일의 개체를 선택한 후, 선택된 개체의 유일변수 관측값 Z를 수령자 파일의 해당 개체에 추가한다.

최근접이웃기법은 데이터 통합과정에서 공통변수 X의 정보만을 이용하나, 회귀분석 기법은 공통변수 X와 제공자 파일의 유일변수 Z를 이용하기 때문에 Ingram et al.(2000)은 실제로 현실에서 데이터 통합 방법에 회귀분석과 같은 예측평균매칭(predicted mean matching)기법이 좋은 성능을 나타낸다고 하였다.

2.3. 수정된 데이터 통합 알고리즘

정성석 등(2004)이 제안한 수정된 데이터 통합 기법은 회귀분석 기법이 예측치의 거리가 가장 가까운 하나의 개체만을 사용함으로써 상대적으로 유사한 개체에 대한 정보가 손실되는 단점을 보완하여 데이터 통합의 성능을 높이고자 회귀분석기법에 k-최근접이웃기

법을 결합하여 k개의 개체를 이용하여 통합변수를 추가시키는 방법이다. 이 방법을 자세히 살펴보면 다음과 같다.

제공자 파일을 대상으로 공통변수 X를 설명변수로 하고 유일변수 Z를 반응변수로 하는 회귀모형을 추정한 후, 제공자 파일에서 추정된 회귀모형을 수령자 파일과 제공자 파일에 적용하여 각 파일에서 유일변수 Z의 예측값 $\hat{Z}$를 구한다. 두 파일에서의 예측값을 이용하여 수령자 파일의 각 개체에 대해 제공자 파일의 모든 개체 사이의 거리를 계산하여 거리가 가장 가까운 k개의 제공자 파일의 개체를 선택한 후, 선택된 개체의 유일변수 관측값 Z를 수령자 파일의 개체에 추가한다. 이때 유일변수가 연속형인 경우 k개 Z값의 평균을, 범주형이면 k개 Z값의 최빈값를 이용한다.

2.4. 군집화 데이터 통합기법

김순영 등(2005)이 제안한 군집화 데이터 통합기법은 데이터 통합과정을 수행하기 전에 제공자 파일을 대상으로 비계층적 군집화 방법의 하나인 c-평균 군집화(일반적으로는 k-means clustering이라 표현하지만 본 논문에서는 앞의 k-NN방법의 k와 혼동을 피하기 위해 k대신 c를 사용한다.)를 적용하여 유사한 개체끼리 몇몇의 집단으로 그룹화시킨 후, 수령자 파일의 개체를 제공자 파일의 군집 중 가장 가까운 군집으로 할당하여 각 그룹별로 통합과정을 수행하는 방법이다. 이 방법을 자세히 살펴보면 다음과 같다.

제공자 파일을 이용하여 데이터를 몇 개의 유사한 그룹으로 나눈 것인지 군집의 수(c)를 결정한 후, 제공자 파일에서 공통변수 X를 이용하여 c-평균 군집화(c-means clustering)과정을 수행한다. 제공자 파일에서 분리된 c개의 각 군집의 중심과 수령자 파일의 각 개체간의 거리를 구하여 수령자 파일의 각 개체를 거리가 가장 가까운 제공자 파일의 군집으로 할당하여 모든 수령자 파일의 개체를 c개의 군집에 할당한다. 각 군집별로 제공자 파일의 유일변수 Z를 반응변수로 공통변수 X를 설명변수로 하여 회귀모형을 추정한 후, 추정된 회귀모형을 해당 군집의 수령자 파일과 제공자 파일에 적용하여 각 파일에서 유일변수 Z의 예측값 $\hat{Z}$를 구한다. 두 파일에서 구한 유일변수 Z의 예측값을 이용하여 수령자 파일의 각 개체에 대하여 예측값 사이에 거리가 가장 가까운 k개의 제공자 파일의 개체를 선택하고, 각 군집에서 선택된 제공자 파일의 k개의 개체들의 유일변수 관측값 Z를 수령자 파일의 해당 개체에 추가한다. 이때, 유일변수가 연속형이면 k개 Z값의 평균을, 범주형이면 k개 Z값의 최빈값을 이용한다.

2.5. 의사결정 규칙을 이용한 데이터 통합기법

본 연구에서 제시한 의사결정 규칙을 이용한 데이터 통합 기법은 데이터 마이닝 기법 중 자료로부터 의사결정 규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 방법인 의사결정나무(classification and regression tree)를 데이터 통합에 응용한 방법으로 기존의 거리 개념을 이용하지 않고 예측의 개념을 이용한다.

제공자 파일에서 의사결정나무를 생성하여 의사결정규칙을 생성한 후, 생성된 의사결정규칙에 수령자 파일의 공통변수를 적용하여 분류 또는 예측된 값을 수령자 파일에 추가

될 통합변수의 예측값으로 한다. 이 방법을 자세히 살펴보면 다음과 같다.

step 1. 제공자 파일에서 유일변수 Z중 임의의 s번째 변수를 목표변수로 공통변수 X를 입력변수로 하여 의사결정나무(decision tree)를 생성한 후 이로부터 의사결정규칙을 도출한다. 이때, s번째 유일변수가 연속형인 경우는 회귀나무(regression tree), 범주형인 경우는 분류나무(classification tree)가 생성된다.

step 2. step1에서 도출된 의사결정 규칙에 수령자 파일의 공통변수 X를 대입하여 수령자 파일에 추가될 s번째 유일변수 Z_s의 예측값 $\hat{Z}_s$를 구한다.

step 3. step2에서 구해진 예측값 $\hat{Z}_s$를 s번째 통합변수 Z'의 값으로 사용 하여 수령자 파일에 추가 한다.

step 4. $s(s = 1, 2, \cdots, S)$번째 유일변수에 대하여 step1에서 step3까지를 반복적으로 적용하여 하나의 통합된 파일을 형성한다.

기존의 대부분의 데이터 통합에 관한 연구는 통합될 변수가 양적 자료를 대상으로 통합시키는 기법에 한정되어있지만 제안하고자 하는 기법은 통합될 변수가 양적 자료뿐만 아니라 질적 자료도 적용 가능한 기법으로 데이터의 측도에 관계없이 데이터 통합을 일괄적으로 적용시킬 수 있다.

3. 사례를 통한 데이터 통합 알고리즘의 비교

기존의 방법과 제안된 방법의 데이터 통합 성능을 비교하기 위해 UCI Repository(Blake and Merz, 1998)에 있는 데이터를 이용하여 모의실험을 하였다.

사례연구에 사용된 데이터는 Abalone, CMC(Contraceptive Method Choice), Letter Recognition, Pen-Based Recognition of Handwritten Digit 그리고 Adult 데이터이고, 이를 이용하여 정성석 등(2004)이 제안한 수정된 데이터 통합 알고리즘, 군집화 데이터 통합 알고리즘과 본 연구에서 제안하는 의사결정규칙을 이용한 데이터 통합 알고리즘의 효율성을 비교하였다. 이때, 수정된 데이터 통합 알고리즘과 군집화 데이터 통합 알고리즘의 성능은 통합에 사용될 개체수(k)를 1, 3, 5, 7로 증가시키며 정확성 측도의 값을 측정하였다.

3.1. 데이터 실험 과정

데이터 통합을 위해서는 별도의 두 개의 데이터 파일이 존재해야 하는데, 여기서는 통합의 정확성을 측정하기 위해 하나의 데이터를 파티션하여 사용하였다.

데이터 파티션은 수령자 파일과 제공자 파일이 포함할 변수의 분리와 개체의 분리과정을 포함하는데, 이는 특별한 법칙이 없어 주관적일 수 있으므로 일정한 규칙을 정한 상태에서 비교실험을 수행하였다. 수령자 파일과 제공자 파일에 포함할 데이터의 비율은 Yoshizoe and Araki(1999)에서 사용한 6대 4로 하였으며, 데이터의 분리는 단순임의(simple random) 방법을 사용하였다. 그리고 각 파일이 포함될 변수의 분리는 데이터 통합에서 가장 중요하게 여겨지는 조건부독립성을 근거로 하였다. 즉, 조건부 독립성이 만족되어야 통계적 매칭이 유용하므로, 이를 역으로 조건부 독립성이 유지되도록 변수를 분리하였다. Rässler(2002)는

이를 회귀분석 접근법(regression approach)으로 판단하는 방법을 제시하였다. 각 파일이 포함할 변수를 분리하기 전에 최종 분석의 목표 변수(target variable)는 데이터 통합에 영향을 주지도 받지도 않도록 하기 위해 수령자 파일의 유일변수 Y에 포함시키기로 하고 회귀분석 접근법을 적용시키면 다음과 같다.

$Z = \beta_0 + \beta_{ZX,Y}X + \beta_{ZY,X}Y$에서 $\beta_{ZY,X} = 0$이면 $\rho_{ZY|X} = 0$으로부터 제공자 파일의 유일변수 Z는 유의수준 0.05에서 목표변수가 설명변수로 유의하지 않은 반응변수이고, 공통변수 X는 제공자 파일의 유일변수로 선택된 변수를 반응변수로 하여 유의한 설명변수를 선택한다. 그리고 제공자 파일의 유일변수와 공통변수에 포함되지 않는 변수를 수령자 파일의 유일변수 Y에 포함시킨다. 각 데이터의 파티션 결과는 표 3.1과 같다.

표 3.1: 실험 데이터의 파티션 결과

데이터	변수		개체수			제공자 파일 유일변수(Z)	공통변수(X)	수령자 파일 유일변수(Y)	군집수
	연속	범주	전체	수령자 파일	제공자 파일				
Abalone	7	1	4,177	2,506	1,671	Length	Diameter Sweight Vweight	Sex, Height Wweight **Rings**	2
CMC	2	8	1,473	884	589	H-edu Nchild Media	W-age, W-edu W-relig, W-work Slindex	H-occup **Cmethod**	2
Letter Recognition	16	1	20,000	12,000	8,000	X-box, Y-box Width, High Onpix, X2Ybar XY2br, Yegvx	X-bar, Y-bar X2bar, Y2bar XYbar, X-ege Xegvy, Y-ege	**Lettr**	4
Handwritten Digits*	16	1	10,992	6,595	4,397	a1, a3, a6 a8, a13	a2, a4, a5, a7 a9, a10, a11 a12, a14 a15, a16	**a17**	3
Adult	6	8	45,222	30,162	15,060	Martial-status Occupation	Age, Sex Education	$o.w$**	·

*: Pen-Based Recognition of Handwritten Digit

**: 제공자 파일의 유일변수와 공통변수를 제외한 나머지 변수들

CMC 데이터의 경우 통합에 사용될 변수 중 Media변수는 이항 범주형 변수이며, Adult 데이터의 통합에 사용될 Martial-status변수는 7범주, Occupation 변수는 14범주를 갖는 범주형 변수이다. 나머지 통합에 사용될 변수는 모두 연속형 변수이다.

파티션된 데이터인 수령자 파일과 제공자 파일을 이용하여 기존의 통합방법과 제안하는 방법을 이용하여 데이터 통합과정을 수행한다.

특히, 기존의 방법은 연속형 변수와 이항범주인 범주형 변수에 적용 가능한 기법이므로 통합될 변수가 모두 다범주 범주형 변수로 구성된 Adult 데이터는 기존의 기법과 비교하지 않고, 제안하는 의사결정 규칙을 이용한 통합과정만 수행하였다.

기존의 군집화 데이터 통합 알고리즘을 사용하기 위해서는 다음과 같은 사항들을 고려하였다. 파티션된 데이터 중 제공자 파일을 c-평균 군집화를 이용하여 유사한 개체끼리 군집을 형성하기 위해, 군집의 수는 다변량 통계분석 기법 중 주성분 분석을 이용하여 고유값의 크기가 1 이상인 주성분의 개수를 군집의 수로 결정하여 사용한다. 회귀모형으로 구

해진 두 파일의 예측치간 거리가 1 이하인 개체만을 통합에 고려하기 위해 회귀모형의 반응변수가 될 제공자 파일의 유일변수 Z가 연속형인 경우는 표준화하여 사용하였다. 통합에 사용될 회귀모형에 설명력 있는 공통변수만 포함되도록 단계적(stepwise) 변수선택을 수행하였다. 또한 두 개체간의 유사성의 척도는 연속형 변수의 경우 유클리디안 거리를 사용하여 수령자 파일의 각 개체에 대해 예측치의 차이가 상대적으로 적은 제공자 파일의 k개(k=1,3,5,7)의 개체를 이용하여 데이터 통합을 수행한다. 각 군집별로 통합과정을 수행한 후 각 군집을 결합하여 최종 통합된 파일을 만든다.

의사결정 규칙을 이용한 통합 알고리즘은 특별히 고려해야 할 조건 없이 2.5절에서 언급한 통합과정을 수행하여 최종 통합된 파일을 만든다.

마지막으로, 통합의 정확성을 평가하기 위한 측도로는 통합변수가 연속형이면 실제값과 통합 알고리즘을 통하여 추가된 값의 차를 이용한 평균제곱오차(MSE), 통합변수가 범주형이면 실제값과 통합 알고리즘을 통하여 추가된 값의 불일치도를 이용한 오분류율을 사용하였다. 정확성의 측도가 작은 통합 알고리즘이 더 효율적인 알고리즘이다.

3.2. 통합 결과 비교

3.1절의 실험과정에서 설명한 과정에 따라 Abalone, CMC, Letter Recognition, Handwritten Digits 그리고 Adult 데이터에 대해 반복실험을 통해 정확성 측도(MSE, 오분류율)를 구하였다. 데이터의 파티션에서 개체의 분리를 단순임의방법을 사용하였으므로, 실험의 반복을 통하여 더 신뢰있는 결론을 내리고자 정성석 등(2004)와 김순영 등(2005)에서 사용한 20회의 반복횟수를 사용하였다. 20회의 반복 실험을 통해 구해진 정확성 측도의 평균을 이용하여 제안된 의사결정 규칙을 이용한 통합 알고리즘과 정성석 등(2004)의 수정된 통합 알고리즘 및 김순영 등(2005)의 군집화 통합 알고리즘의 성능을 비교하였다.

통합과정의 반복실험을 통한 데이터 통합의 정확도를 평가한 결과는 표 3.2와 같고, 그림 3.1, 그림 3.2, 그림 3.3, 그림 3.4 그리고 그림 3.5를 통해 확인할 수 있다.

데이터 질을 향상시키기 위해 기존의 통합기법과 제안한 통합기법을 비교해본 결과 기존의 통합기법은 통합에 사용하는 개체의 수 k가 1에서 7까지 증가할수록 두 방법 모두 정확성의 측도인 MSE와 오분류율이 점차 감소하나 제한하는 의사결정 규칙을 이용한 통합기법의 정확성이 훨씬 작은 값을 취하므로, 제한하는 통합기법이 더 정확한 데이터 통합을 수행함을 알 수 있었다.

통합 변수가 연속형 변수의 경우 다른 기법에 비해 보다 정확한 데이터 통합작업을 수행함을 알 수 있었으며, 범주형 변수일 경우는 이항 범주형 변수일 경우는 제한하는 방법이 보다 정확한 데이터 통합 작업을 수행하나, 범주의 수가 많은 다범주 변수의 경우는 통합 작업이 범주의 수에 민감하여 범주의 수가 많을수록 오분류율이 급격히 증가함을 알 수 있었다.

의사결정 규칙 통합 알고리즘과 기존의 통합 알고리즘에 의한 결과를 살펴보면 다음과 같다.

Letter Recognition 데이터의 경우 모든 변수에 대하여 의사결정 규칙을 이용한 데이터 통합 알고리즘의 MSE가 더 작음을 알 수 있다.

Handwritten Digit 데이터의 경우 모든 변수에 대하여 의사결정 규칙을 이용한 데이터 통합 알고리즘의 MSE가 월등히 작음을 알 수 있다. 즉, 의사결정 규칙을 이용한 통합 알고리즘이 더 정확한 데이터 통합을 수행한다.

Abalone 데이터의 경우 k가 5 또는 7로 증가할 경우 의사결정 규칙을 이용한 알고리즘보다 기존의 알고리즘의 MSE가 더 작음을 알 수 있었으나, 거의 비슷한 정확도를 나타냄을 알 수 있다.

CMC 데이터의 경우 모든 변수에 대하여 의사결정 규칙을 이용한 데이터 통합 알고리즘의 MSE가 더 작음을 알 수 있다. 즉, 의사결정 규칙을 이용한 통합 알고리즘이 더 정확한 데이터 통합을 수행한다.

Adult 데이터의 경우 통합될 변수가 다범주를 갖는 범주형 변수이므로 기존의 통합 기법은 적용하지 않고, 의사결정 규칙을 이용한 통합 알고리즘만 적용의 결과 범주의 수에 민감하여 범주의 수가 늘어남에 따라 오분류율이 급격하게 늘어남을 알 수 있었다.

표 3.2: 데이터 실험 결과

데이터명	변수명	MSE or 오분류율(%)				
		k / 통합기법	1	3	5	7
CMC	H-edu	reg+knn	0.80068	0.53185	0.48238	0.46223
		Cluster	0.76193	0.52144	0.47628	0.45661
		Tree	0.42675			
	Nchild	reg+knn	7.11810	4.76964	4.30438	4.12376
		Cluster	7.22342	4.79493	4.28528	4.11323
		Tree	4.06003			
	Media	reg+knn	10.78600	9.12400	8.67050	8.37750
		Cluster	11.26650	8.82850	8.51150	8.25050
		Tree	7.66950			
Abalone	a1	reg+knn	0.00063	0.00043	0.00039	0.00038
		Cluster	0.00061	0.00042	0.00038	0.00037
		Tree	0.00042			
Handwritten Digits	a1	reg+knn	1114.874	746.078	671.648	640.587
		Cluster	898.964	603.737	543.153	517.729
		Tree	375.125			
	a3	reg+knn	575.477	383.154	344.225	327.571
		Cluster	508.114	339.780	306.107	291.684
		Tree	293.177			
	a6	reg+knn	240.224	160.184	144.021	136.913
		Cluster	165.330	111.268	100.768	95.877
		Tree	73.241			

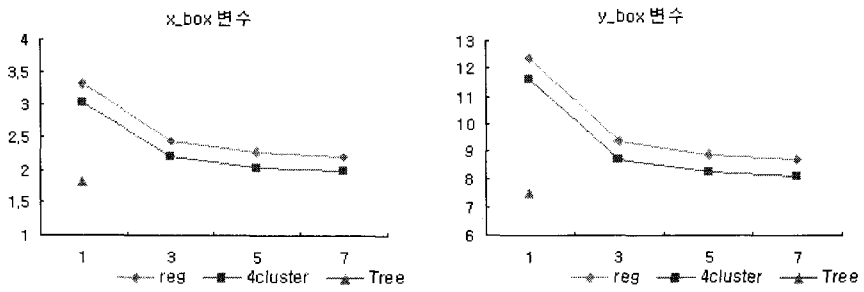


그림 3.1: Letter Recognition 데이터의 MSE

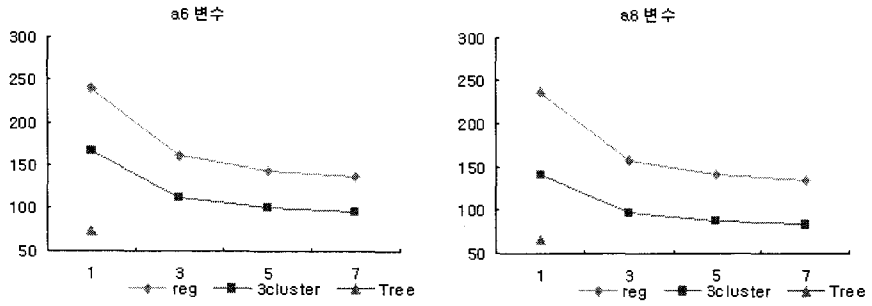


그림 3.2: Handwritten Digits 데이터의 MSE

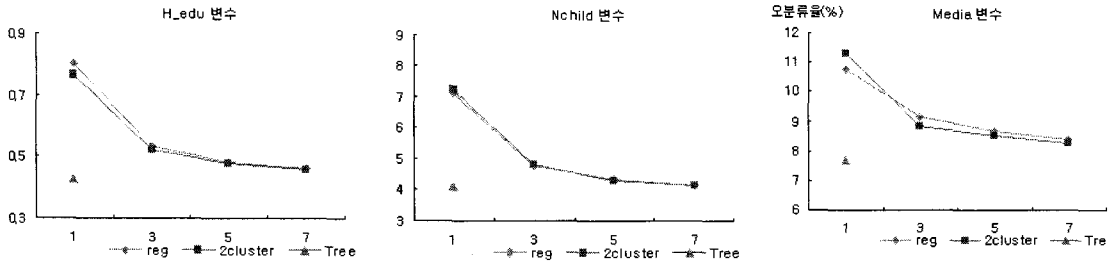
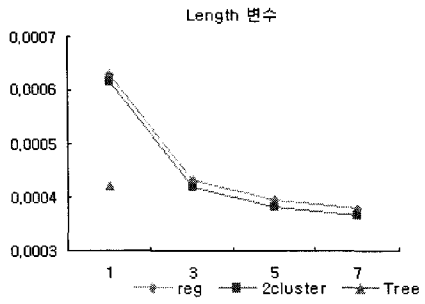


그림 3.3: CMC 데이터의 MSE 및 오분류율(%)

표 3.2 계속 : 데이터 실험 결과

데이터명	변수명	MSE or 오분류율(%)				
		k / 통합기법	1	3	5	7
Handwritten Digits*	a8	reg+knn	236.354	158.098	141.800	134.876
		Cluster	141.325	96.575	87.162	83.043
		Tree	67.379			
	a13	reg+knn	399.572	263.901	237.589	226.635
		Cluster	281.934	191.325	173.716	165.862
		Tree	161.206			
Letter Recognition	x-box	reg+knn	3.32141	2.43658	2.27565	2.20887
		Cluster	3.03555	2.20894	2.04659	1.98200
		Tree	1.80944			
	y-box	reg+knn	12.37523	9.35064	8.89108	8.73279
		Cluster	11.62294	8.73658	8.27886	8.12529
		Tree	7.47849			
	width	reg+knn	3.07275	2.32746	2.20523	2.16922
		Cluster	2.99889	2.21389	2.08219	2.03548
		Tree	1.97525			
	high	reg+knn	5.31536	4.11682	3.97210	3.94985
		Cluster	5.32382	4.05162	3.87330	3.82841
		Tree	3.67595			
	onpix	reg+knn	2.72629	2.04614	1.94107	1.90977
		Cluster	2.81865	2.06460	1.91978	1.86180
		Tree	1.76962			
	x2ybr	reg+knn	4.26628	3.45709	3.35354	3.33558
		Cluster	2.89559	2.25587	2.15821	2.11776
		Tree	1.70635			
	xy2br	reg+knn	4.42424	3.53794	3.40872	3.39133
		Cluster	3.36229	2.60125	2.46880	1.99142
		Tree	2.02089			
	yebvx	reg+knn	3.44726	2.56054	2.39379	2.32045
		Cluster	2.88942	2.13041	2.41729	1.93350
		Tree	1.72731			
Adult	Martial-status	Tree	7범주	32.65		
	Occupation	Tree	14범주	68.82		

* : Pen-Based Recognition of Handwritten Digit
reg+knn : 정성석 등(2004)의 수정된 통합 알고리즘
Cluster : 제안하는 군집화 통합 알고리즘
reg+knn : 의사결정 규칙을 이용한 통합 알고리즘



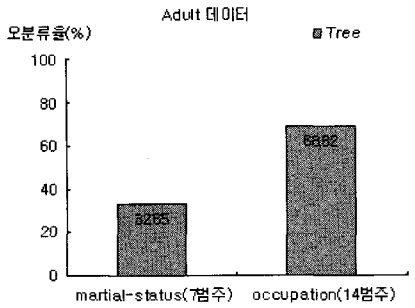


그림 3.4: Abalone 데이터의 MSE 그림 3.5: Adult 데이터의 오분류율(%)

4. 결론 및 향후 연구과제

본 연구에서는 마이닝에 사용될 데이터의 질을 향상시키기 위한 방법으로 데이터 보강 단계에서 사용할 수 있는 방법인 데이터 통합 기법을 살펴보고, 마이닝 기법으로 사용되고 있는 기계학습 기법 중 의사결정 규칙을 이용한 데이터 통합 알고리즘을 제시하였다.

실험 결과 정성석 등(2004)이 제안하였던 기존의 회귀분석기법에 k-최근접이웃기법을 적용한 방법과 김순영 등(2005)이 제안하였던 군집화 기법을 이용한 데이터 통합기법 보다 본 연구에서 제안한 의사결정 규칙을 이용한 데이터 통합기법이 더 정확한 데이터 통합 작업을 수행함을 알 수 있었다. 연속형 변수와 범주의 수가 적은 범주형 변수의 경우는 제안하는 방법이 좀더 효율적인 데이터 통합작업을 수행하나, 범주의 수가 많은 범주형 변수의 경우는 범주의 수가 증가함에 따라 오분류율이 급격히 증가함을 알 수 있었다.

향후 연구 과제로는 다범주 범주형 변수의 통합에 관한 연구가 이루어져야 할 것이다. 의사결정 규칙을 이용함으로써 통합될 변수의 측도에 관계없이 다범주 범주형 변수의 통합도 가능하게 되었지만, 범주의 수에 민감하다는 문제점을 지내고 있다. 또한 대부분 상당수의 범주형 변수들이 3개 이상의 범주를 가지므로 이러한 경우 효율적 데이터 통합작업을 수행할 수 있는 다른 통계적 기법 및 마이닝 기법에 관한 연구도 가치가 있을 것이다.

참고문헌

김순영, 이기훈, 정성석 (2005), A Study on a Statistical Matching Method Using Clustering for Data Enrichment, <한국통계학회 논문집>, **12**, 509-520.

정성석, 김순영, 김현진 (2004). 데이터 보강을 위한 데이터 통합기법에 관한 연구, <응용통계연구>, **17**, 605-617.

Ingram, D., O' Hare, J., Scheuren, F. and Turek, J. (2000). Statistical matching: a new validation case study, *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Rässler, S. (2002). *Statistical Matching : A frequentist theory, practical applications, and alternative Bayesian approaches*, Springer Verlag, New York.

Saporta, G. (2002). Data fusion and data grafting, *Computational Statistics & Data Analysis*, **38**, 465-473.

U.S. Department of Commerce, (1980). Report on exact and statistical matching techniques, *Statistical Policy Working Paper 5. Washington, DC: Federal Committee on Statistical Methodology.*

van der Putten, P., Joost N. K. and Gupta, A. (2002). Why the Information explosion can be bad for data mining, and how data fusion provides a way out, *Second SIAM International Conference on Data Mining*, Arlington, April, 11-13.

Yoshizoe, Y. and Araki, M. (1999). Use of statistical matching for household surveys in Japan, *In 52nd Session of the International Statistical Institute*, Helsinki, Finland.

[2006년 1월 접수, 2006년 3월 채택]

A Study on the Data Fusion Method using Decision Rule for Data Enrichment*

S. Y. Kim[1] S. S. Chung[2]

ABSTRACT

Data mining is the work to extract information from existing data file. So, the one of best important thing in data mining process is the quality of data to be used. In this thesis, we propose the data fusion technique using decision rule for data enrichment that one phase to improve data quality in KDD process. Simulations were performed to compare the proposed data fusion technique with the existing techniques. As a result, our data fusion technique using decision rule is characterized with low MSE or misclassification rate in fusion variables.

Keywords: Data fusion, Statistical matching, Data enrichment, Data Mining, k-Nearest Neighbor, Decision Tree, Recipient file, Donor file.

* This research was supported by the Ministry of Commerce, Industry and Energy Development through Core Technology Development Program(105077001).

1) Doctoral Student, Department of Statistical Informatics, Chonbuk National University, 664-14 1 ga Duckjin-Dong Duckjin-Gu Chonju Chonbuk, 561-756, Korea.
E-mail: rabbit@chonbuk.ac.kr

2) (Corresponding author) Professor, Department of Statistical Informatics, Chonbuk National University, 664-14 1 ga Duckjin-Dong Duckjin-Gu Chonju Chonbuk, 561-756, Korea.
E-mail: sschung@chonbuk.ac.kr