응용통계연구 제 17권 3호, 2004년, pp. 605-617

데이터 보강을 위한 데이터 통합기법에 관한 연구*

정성석[1] 김순영[2] 김현진[3]

요 약

데이터마이닝에서 가장 중요한 요소 중 하나는 마이닝에 사용될 데이터의 질이다. 질 높은 데이터를 바탕으로 마이닝이 수행될 때, 데이터마이닝의 잠재적 가치는 증대될 것이다. 본 논문에서는 지식발견 과정 중 데이터의 질을 향상시키기 위한 한 단계인 데이터 보강을 위해 데이터 통합 기법을 제안하고, 모의실험을 통해 제안된 알고리즘의 효율성을 비교하였다. 실험결과 제안된 알고리즘이 데이터 통합의 성능을 향상시킴을 알 수 있었다.

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

1. 서론

요즘 우리는 무분별한 직접우편(Direct Mail)과 텔레마케팅(Tele-Marketing)의 홍수 속에서 불편함을 느끼고 있지만, 이를 시도하는 기업은 여러 가지 심각한 문제에 직면하고 있다. 이 중 하나는 사전 정보가 충분하지 못한 상태에서 발송된 우편이나 캠페인성 전화가 해당 기업의 물품구매에 가능성이 높은 고객에게 전달되지 못해 그 효과를 상실하게 되고 고객과의 지속적인 관계형성에서 나쁜 이미지를 심어 주게 된다. 이러한 문제의 근본적인 해결방안은 기업에서 다양한 경로를 통해 축적한 다량의 데이터로부터 기업의 의사결정에 도움을 줄 수 있는 정보를 추출하기 위해 데이터를 효과적으로 분석하는 데서 찾을 수 있다. 이런 고객에 대한 축적된 데이터를 이용하여 많은 기업이 시도하고 있는 대표적인 분석적 접근 중 하나가 데이터마이닝(Data Mining)이다. 데이터마이닝은 국내 일부 기업 등에 도입되어 필요성과 효과에 대한 공감대가 형성되고 있으나 실제 진행 과정에서는 여러 문제점이 나타나고 있다. 그 중 핵심적인 사항이 데이터 수집 문제이다. 데이터 분석을 통해 고객에게 접근하기 위해서는 우리 기업의 주요 고객은 어떤 사람들이고, 어떠한 고객이 우리에게 가장 많은 가치를 줄 수 있는가에 대한 다양한 데이터가 필요하다. 그러나 현재 활용되고 있는 대부분의 데이터는 기업활동으로부터 부수적으로 생성된 고객들의 이용실

* 본 연구는 과학기술부 주관 인간기능생활지원로봇기술개발사업의 지원에 의해 이루어졌음.

1) (561-756) 전라북도 전주시 덕진구 덕진동 1가 664-14, 전북대학교 수학통계정보과학과 교수
E-mail: sschung@chonbuk.ac.kr

2) (561-756) 전라북도 전주시 덕진구 덕진동 1가 664-14, 전북대학교 통계정보과학과 박사과정
E-mail: rabbit@chonbuk.ac.kr

3) (561-756) 전라북도 전주시 덕진구 덕진동 1가 664-14, 전북대학교 통계정보과학과 석사과정 졸업
E-mail: jinijoa98@hanmail.com

적 위주의 데이터만 존재할 뿐 고객을 설명하고 이에 대한 접근이 가능한 데이터는 거의 없는 실정이다. 이러한 데이터 수집에 대한 어려움을 데이터 보강(Data Enrichment)에 의해서 해결 할 수 있다. 이 데이터보강을 위하여 데이터 통합(Data Fusion) 기법을 사용할 수 있다. 데이터 통합 기법에 관한 기존 연구들은 거리와 같은 유사성 측도를 이용하여 가장 유사한 개체(Nearest Neighbor)를 찾거나, 회귀분석(Regression)과 군집분석(Clustering) 등의 마이닝기법을 적용한 데이터 통합 기법이 제안되었다. 이 중 회귀분석방법에서는 추정치의 거리가 가장 가까운 하나의 개체만을 사용함으로써 상대적으로 유사한 다른 개체들의 정보를 무시하게 되어 더 정확한 데이터 통합을 이룰 기회를 상실하게 된다.

이러한 문제점을 보완하고자 본 연구에서는 회귀분석기법에 k-최근접이웃 기법을 적용하여 상대적으로 유사한 개체에 대한 정보 손실을 줄임으로써 데이터 통합기법의 성능을 높이고자 하였다. 그리고 제안된 데이터 통합 알고리즘과 기존의 알고리즘의 성능을 비교하기 위해 실제 데이터를 이용하여 데이터 통합을 수행한 결과 연속형 변수에 대해 통합이 수행 될 때 제안된 통합기법이 보다 정확한 작업을 수행함을 알 수 있었다.

2절에서는 데이터 보강을 위한 데이터통합의 개념과 데이터 통합의 원류(Origin)를 찾기 위해 매칭(Matching)에 대해 알아보고, 3절에서는 기존의 데이터통합기법에 대해 알아본 후 수정된 데이터통합기법을 제안하였다. 4절에서는 제안한 방법을 실제 데이터에 적용해보고 그 성능을 기존의 방법과 비교하였다. 마지막으로 5절에서는 결론과 향후 연구 방향에 대해 논의하였다.

2. 데이터 보강과 데이터 통합

데이터 보강은 분석하고자 하는 데이터에 기존 정보를 결합하여 그 양(개체 수)과 깊이(변수의 수)를 늘리는 것이라 정의할 수 있으며, 이를 위해 데이터를 구입하거나 고객과의 인터뷰를 통해 얻은 자료를 기존의 정보와 결합한다. 데이터 보강을 통해 데이터의 충실도를 상당히 높일 수 있는데, 여기서 데이터의 충실도란 데이터의 정확도, 데이터의 양, 데이터의 깊이에 의해 평가된다.

데이터 보강을 위한 방법 중의 하나인 데이터 통합은 보유하고 있는 데이터 파일에 필요한 변수가 없거나, 결측치가 존재할 경우 다른 원천 데이터로부터 모은 자료와 정보(information)를 통합시키는 것이라고 정의한다(Saporta, 2002).

데이터 통합은 대용량 데이터의 병합(macro data set merging), 통계적 개체 연결(statistical record linkage) 또는 다원천 대체(multi-source imputation)로 알려져 있고 오늘날까지 조사(survey)에서 응답자 수와 설문 문항 수를 줄이는데 이용되고 있다. 예를 들어, 상품과 미디어에 관한 Belgian National Readership의 조사는 각 10,000명을 포함하는 다른 두 그룹의 응답자를 조사하여, 하나의 조사자료로 통합하였다. 이를 통해 각각의 응답자에게 들이는 시간과 비용을 줄일 수 있었다. 그리고 추가된 변수는 일반적으로 예측의 질을 향상시킨다(van der Putten et al., 2002).

데이터 통합을 구체적으로 살펴본 모형도는 그림 2.1과 같다. 수령자 파일(Recipient file)은 앞으로 통합되어질 데이터 파일이고, 제공자 파일(Donor file)은 통합을 위해 추가

적인 정보를 제공하기 위해 사용될 데이터 파일이다. 이 두 데이터 파일에 존재하는 공통변수(Common Variables)를 X로 표시하고, 각각의 파일에 유일하게 존재하는 유일변수(unique variables)를 수령자 파일에서는 Y, 제공자 파일에서는 Z로 표시한다.

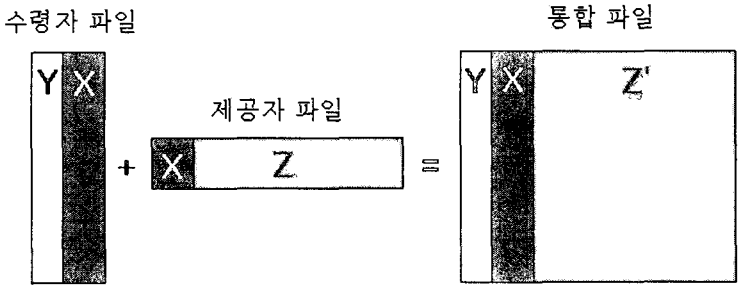

그림 2.1: 데이터 통합 개념도

데이터 통합 과정(Data Fusion Procedure)은 공통변수 X를 이용하여 제공자 파일에만 존재하는 유일변수 Z를 수령자 파일에 추가하여 통합된 파일(fused file)을 완성하는 과정이다. 이 통합과정을 통해 수령자 파일에 추가된 변수들을 Z'로 표시하고, 통합변수(fusion variables)라고 한다.

데이터 통합은 통계적 매칭(Statistical Matching)과 같은 의미로 사용된다. 여기서 매칭이란 같은 모집단의 개체가 포함되어 있는 두개 또는 그 이상의 데이터 파일로부터 개체를 연결하는 것이라 정의된다. 우리가 통계분석을 수행할 때, 필요로 하는 변수를 모두 포함하는 데이터 파일은 흔하지 않기 때문에 이를 해결하기 위한 방법으로 첫째 필요한 변수를 포함한 데이터를 다시 수집, 둘째 통계적 기법을 사용해서 값을 할당(assign)하거나 대체(imputation), 셋째 여러 데이터 파일을 이용해서 필요한 변수를 매칭(matching)시켜 사용한다. 이중 세 번째에 해당하는 것이 위에서 말하는 매칭방법이고, 매칭은 통계분석을 수행하기 위한 준비단계로서 분석에 필요로 하는 변수를 추가하기 위해 사용될 수 있는 기술 중 하나이다. 매칭을 통한 방법은 다른 조사를 통해서 데이터를 얻는 것보다 시간과 비용을 절약 할 수 있고 때로는 분석과 추정에 있어서 더욱 신뢰성을 높이는 방법일 수 있으며, 조사 응답자의 부담을 줄여줄 수 있다(U.S. Department of Commerce, 1980). 그리고, 매칭은 크게 정확한 매칭(Exact Matching)과 통계적 매칭(Statistical Matching)으로 구분된다.

정확한 매칭은 제공자 파일이 수령자 파일의 모든 개체를 포함하는 경우 사용할 수 있는 방법으로, 서로 다른 데이터 파일로부터 같은(same) 개체를 연결하는 방법을 말하며, Actual Matching 또는 Object Matching이라고 한다. 통계적 매칭은 제공자 파일과 수령자 파일에 공통으로 포함되는 개체가 적거나 없는 경우 사용되는 방법으로 서로 다른 데이터 파일로부터 같은(same) 개체를 연결한다기보다는 유사한(similar) 개체를 연결하는 방법을 말하며, 이를 Synthetic Matching, Stochastic Matching, Attribute Matching 또는 Data Matching이라고 한다.

3. 데이터 통합 기법

일반적인 데이터 통합과정은 초기의 두 파일이 주어지면, 수령자 파일의 한 개체와 제공자 파일의 모든 개체 사이의 거리(distance)를 계산한 후, 그 중 가장 작은 값을 갖는 제공자 파일의 개체를 선택하여 수령자 파일에 추가시키는 것이다. van der Putten et al.(2002)은 데이터 통합 알고리즘이 유용한 결과를 도출하기 위해 다음과 같은 제약조건을 제시하였다. 첫째, 제공자 파일은 수령자 파일을 대표할 수 있어야 한다. 그러나 반드시 두 데이터가 같은 모집단에서 추출될 필요는 없다. 둘째, 공통변수X가 주어졌을 때, 유일변수인 Y와 Z 사이에 조건부 독립관계가 성립되어야 한다. 조건부독립성은 통계적 매칭에서 유용한 가정으로 Rässler(2002)가 제시한 회귀분석접근법(regression approach)으로 판단할 수 있다.

본 절에서는 기존 연구의 통계적 매칭 알고리즘인 k-최근접이웃방법과 회귀분석방법에 대해서 살펴보고, 수정된 데이터 통합기법을 제안하였다.

3.1. k-최근접이웃기법

최근접이웃방법은 통계적 매칭에 가장 흔히 사용되는 방법으로 가장 유사한 하나의 개체를 매칭에 사용하는 방법이다. 여기서 한 단계 나아가 상대적으로 유사한 k개의 개체를 선택하여 매칭에 사용하는 방법이 k-최근접이웃 방법이다. van der Putten et al.(2002)에 의해 제시된 데이터 통합은 공통변수 X를 이용하여 가장 가까운 k개의 개체를 선택한 후, 이를 이용해 통합변수를 추가하는 방식으로 이루어진다. 이 방법을 자세히 살펴보면 다음과 같다.

모든 공통변수 X를 수치형(numerical)으로 변환하고, 이를 이용하여 수령자 파일의 각 개체에 대해 제공자 파일의 모든 개체와의 거리를 계산한다. 거리 계산은 유클리디안 거리(Euclidean distance)를 흔히 사용한다. 계산한 거리 중 수령자 파일의 각 개체와 가장 가까운 제공자 파일의 k개의 개체를 선택한다. 선택된 k개 개체에 해당하는 제공자 파일의 유일변수 Z를 이용하여 수령자 파일의 각 개체에 통합변수를 추가시킨다. 이때, 유일변수가 연속형이면 k개 Z값의 평균(mean)을, 범주형이면 k개 Z값의 최빈값(mode)을 이용한다.

3.2. 회귀분석 기법

회귀분석을 적용하여 매칭을 하는 방법은 먼저 하나의 데이터 파일에서 회귀모형을 추정한 후, 추정된 회귀모형을 이용하여 두개의 데이터 파일에서 예측치를 구한다. 그리고 두 파일의 예측치 사이의 거리가 가장 짧은 개체를 찾음으로써 매칭이 이루어진다. Ingram et al.(2000)에서 제시된 회귀모형 접근방법은 다음과 같다.

제공자 파일의 유일변수 Z중 임의의 s번째 변수를 목표변수로, 제공자 파일의 공통변수 X를 설명변수로 하여 회귀모형을 추정한다. 추정된 회귀모형을 수령자 파일과 제공자 파일에 적용하여 각 파일에서 s번째 유일변수 Z_s의 예측치을 구한다. 두 파일에서의 예측값을 이용하여 수령자 파일의 각 개체에 대해 모든 제공자 파일 개체와의 거리를 구한다. 이를 이용하여 수령자 파일의 각 개체에 가장 가까운 제공자 파일에 해당하는 개체의 유일

변수 Z_s를 수령자 파일의 해당 개체에 추가한다. 이때 주의할 점은 수령자 파일에 추가되는 값은 거리를 계산할 때 사용한 예측값 $\hat{Z}_s$이 아니고 관측값 Z_s라는 것이다.

회귀분석에 의한 데이터 통합 접근방법은 단순히 공통변수의 거리함수를 이용한 최근접이웃 방법과는 다르다. 최근접이웃 접근방법은 데이터 통합이 이루어질 때 공통변수 X만을 이용하지만, 회귀분석 접근방법은 공통변수 X뿐만 아니라 제공자 파일의 유일변수 Z를 이용한다는데 그 차이가 있다. Ingram et al.(2000)은 실제로 현실에서 데이터 통합 접근방법에 회귀분석과 같은 예측평균매칭(predicted mean matching)기법은 좋은 성능을 나타낸다고 하였다.

3.3. 수정된 데이터 통합 기법

회귀분석 방법을 이용한 통계적 매칭방법은 추정치의 거리가 가장 가까운 하나의 개체만을 사용함으로써 상대적으로 유사한 다른 개체들의 정보를 무시하게 된다. 본 연구에서 제시한 수정된 데이터 통합기법은 상대적으로 유사한 개체에 대한 정보 손실을 줄여 데이터 통합기법의 성능을 높이고자 회귀분석기법에 k-최근접이웃 접근법을 결합하여 가장 가까운 하나의 개체가 아니라 k개의 개체를 이용하여 통합변수를 추가시키는 방법이다. 이 방법을 자세히 살펴보면 다음과 같다.

step 1. 제공자 파일에서 유일변수 Z중 임의의 s번째 변수를 목표변수로 공통변수 X를 설명변수로 하여 회귀모형을 추정한다.

step 2. 추정된 회귀모형을 수령자 파일과 제공자 파일에 적용하여 각 파일에서 s번째 유일변수 Z_s의 예측치를 구한다.

step 3. 두 파일에서의 예측값을 이용하여 수령자 파일의 각 개체에 대해 모든 제공자 파일의 개체와의 거리를 구한다.

step 4. 계산한 거리를 이용하여 수령자 파일의 각 개체에 가장 가까운 제공자 파일에 해당하는 k개의 개체를 선택한다.

step 5. 선택된 제공자 파일의 k개 개체들의 유일변수 Z_s들의 평균이나 최빈값을 구한 후 이 값을 수령자 파일의 해당 개체에 추가한다. 이때, 유일변수가 연속형이면 k개 Z값의 평균(mean)을, 범주형이면 k개 Z값의 최빈값(mode)을 이용한다.

4. 사례연구

실제 데이터를 이용하여 제안된 방법과 기존의 방법의 데이터 통합의 성능을 비교하였다.

사례연구에 사용된 데이터는 UCI 데이터베이스(http://kdd.ics.uci.edu/)의 데이터 이며, 보다 일반적인 결론을 내리기 위해 Housing, Abalone, CMC (Contraceptive Method Choice), Letter Recognition와 같은 여러 데이터를 사용하여 실험을 수행하였다. 본 연구에서는 k가 1일 때보다 3, 5, 7로 증가함에 따라 정확성 측도의 값을 비교하여 새롭게 제안한 데이터 통합 알고리즘의 효율성을 증명해 보이고자 한다.

4.1. 실험 설계

데이터 통합은 통합시키고자하는 두 개의 데이터가 있다는 가정 하에 가능하나 본 연구에서는 제안된 알고리즘의 성능을 기존 알고리즘과 비교하기 위해 하나의 데이터를 분리하여 사용하였다.

실험과정은 먼저 하나의 데이터를 파티션하여 수령자 파일과 제공자 파일로 분리한 후 기존의 데이터 통합 알고리즘과 이를 개선시키고자 제안한 알고리즘을 사용하여 데이터를 통합한다. 그런 다음 실제값과 통합된 값의 차이를 근거로 두 알고리즘의 성능을 비교하였다.

데이터 파티션 및 회귀모형의 적합과정은 SAS를 사용하였으며, 데이터 통합 알고리즘은 Visual C++로 구현하였다.

4.1.1. 데이터 파티션

실험의 첫 번째 단계인 데이터 파티션은 수령자 파일과 제공자 파일에 포함할 변수와 개체를 분리한다. 수령자 파일과 제공자 파일로 개체분리를 위한 데이터 비율은 Yoshizoe and Araki(1999)에서 사용한 60%대 40%로 하고, 데이터의 분리는 단순임의(simple random)방법을 사용하였다. 각 파일에 포함될 변수 분리는 통계적 매칭에서 가장 중요하게 여겨지는 조건부독립성이 만족되도록, Rässler(2002)가 제시한 회귀분석접근법으로 판단하는 방법을 이용하여 조건부독립성 가정이 유지되도록 변수를 분리하였다.

Housing 데이터를 이용하여 자세히 살펴보면, 각 변수를 회귀분석의 종속변수로 하고 나머지 변수들을 설명변수에 포함시켜 회귀모형을 적합시킨다. 각 파일이 포함할 변수를 분리하기 전에 최종 분석의 목표 변수(target variable)가 될 MEDV는 데이터 통합에 영향이 없도록 하기 위해 수령자 파일의 유일변수 Y에 포함시키기로 하였다. 결국 다음 식

$$Z = \beta_0 + \beta_{ZX,Y}X + \beta_{ZY,X}Y \text{ 에서 } \beta_{ZY,X} = 0 \text{ 이면 } \rho_{ZY|X} = 0$$

으로부터 제공자 파일의 유일변수 Z는 유의수준 0.05하에 MEDV가 설명변수로서 유의하지 않은 반응변수인 INDUS, CHAS, AGE로 선택되었다. 단, 범주형 변수인 CHAS는 유의수준 0.05로 검정하면 선택되지 않으나, 범주형 변수에 대한 통합과정을 설명하기 위해 유의수준 0.01로 검정하여 선택되었다. 그리고 공통변수 X는 제공자 파일의 유일변수로 선택된 INDUS, CHAS, AGE변수를 반응변수로 하여 유의한 설명변수를 선택한다. 여기서, CHAS는 이항변수이므로 로지스틱 회귀모형을 적합시켰다. 이 과정에서 일반적으로 사용하는 0.05나 0.01과 같은 유의수준을 크게 하여 사용하면 공통변수 X에 포함되는 변수가 너무 많아져 데이터 통합이 무의미할 수 있으므로 유의수준이 0.0001이하인 변수만을 공통변수 X로 선택하여 NOX, RM, DIS, RAD, TAX, LSTAT이 선택되었다. 그리고 공통변수에 포함되지 않고 제공자 파일의 유일변수에도 포함되지 않는 변수는 수령자 파일의 유일변수 Z에 포함시킨다. 각 데이터의 파티션 결과는 표 4.1과 같다.

4.1.2. 데이터 통합

수령자 파일과 제공자 파일로 분리된 데이터파일을 통합하는 과정은 근본적으로 회귀

표 4.1: 실험 데이터의 파티션 결과

데이터	변수		개체수			제공자 파일 유일변수(Z)	공통변수(X)	수령자 파일 유일변수(Y)
	연속	범주	전체	수령자 파일	제공자 파일			
Housing	13	1	506	304	202	INDUS CHAS AGE	NOX, RM DIS, RAD TAX, LSTAT	PTRATI CRIM, ZN B, **MEDV**
Abalone	7	1	4,177	2,506	1,671	Length	Diameter Sweight Vweight	Sex, Height Wweight **Rings**
CMC	2	8	1,473	884	589	H-edu Nchild Media	W-age, W-work W-relig, W-edu Slindex	H-occup **Cmethod**
* Letter recognition	16	1	2,000	1,200	800	X-box, Y-box Width, Onpix High, X2Ybar XY2br, Yegvx	X-bar, Y-bar X2bar, Y2bar XYbar, X-ege Xegvy, Y-ege	**Lettr**

* : 개체수가 많고, 선택된 Z변수가 많아 계산량이 많아지므로 원 데이터의 10% 랜덤표본을 뽑아 실험을 수행함.

분석 기법(k=1)을 사용하며 k-최근접이웃 접근법(k=3,5,7)을 적용하였다.

본 연구에서 수행한 데이터 통합과정은 다음과 같은 사항들을 고려하였다. 첫째, 회귀모형으로 구해진 두 파일의 예측치 차이가 1이하인 개체만을 통합에 고려하기 위해 회귀모형의 반응변수가 될 제공자 파일의 유일변수 Z가 연속형인 경우에는 표준화시킨다. 둘째, 제공자 파일에서 회귀모형을 적합할 때, 통합에 사용될 회귀모형에 설명력 있는 공통변수만 포함되도록 단계적(stepwise) 변수선택을 수행한다.

Housing 데이터를 이용하여 변수의 통합을 살펴보면 아래와 같다.

(1) 연속형 변수의 통합

제공자 파일의 연속형 유일 변수인 INDUS의 매칭의 경우, INDUS를 S_INDUS로 회귀분석 전에 표준화한 후, S_INDUS를 반응변수로 하여 회귀모형을 적합시킨 회귀식을 이용하여 제공자 파일뿐만 아니라 수령자 파일에서 S_INDUS의 예측값을 구한 후, 수령자 파일의 하나의 개체에 대해서 202개의 제공자 파일의 모든 개체에 대해 예측값의 차이를 구한다. 그리고 각 수령자 파일의 개체에 대해 예측치의 차이가 상대적으로 작은 제공자 파일의 INDUS를 이용해 매칭이 이루어진다. 본 연구에서는 차이가 가장 작은 1, 3, 5, 7개의 제공자 파일의 개체를 각각 사용해서 매칭을 수행하였다. 표 4.2는 수령자 파일의 각 개체에 대해 예측치의 차이가 작은 것 순으로 7개의 제공자 파일 개체를 보여준다.

표4.2에서 수령자 파일의 첫 번째 개체는 제공자 파일의 16번째 개체와 가장 가까운 예측

표 4.2: 가장 가까운 7개의 예측치의 차이

R	1			···	218			···	304		
k	D	차이	INDUS		D	차이	INDUS		D	차이	INDUS
1	16	0.0090	8.14		143	0.0027	18.10		42	0.0046	8.56
2	10	0.0106	8.14		160	0.0039	18.10		100	0.0826	3.97
3	129	0.0148	9.90		150	0.0042	18.10		45	0.0865	8.56
4	13	0.0248	8.14		156	0.0194	18.10		67	0.0883	19.58
5	9	0.0280	8.14		158	0.0213	18.10		43	0.1029	8.56
6	39	0.0288	3.41		174	0.0219	18.10		125	0.1072	9.90
7	74	0.0303	4.05		165	0.0224	18.10		126	0.1087	9.90

R: 수령자파일의 개체, D: 제공자파일의 개체, INDUS: 제공자파일의 실제값을 나타냄.
차이 $= |S\_\widehat{INDUS}_R - S\_\widehat{INDUS}_D|$

치를 갖는다. 기존의 회귀분석 접근법으로 데이터 통합을 수행한다면 제공자 파일의 16번째 개체의 INDUS값 8.14를 수령자 파일에 추가시킨다. 그러나 제안된 방법에 의해 가장 가까운 3개의 개체를 고려한다면 제공자 파일의 16번째, 10번째, 129번째 개체에 해당하는 INDUS값의 평균 $(8.14+8.14+9.90)/3 = 8.727$을 추가시킨다. 이때 주의할 점은 매칭에 직접 사용되는 값은 예측된 S_INDUS의 값이 아니라 실제 INDUS값을 이용한다는 것이다. 이런 과정을 수령자 파일의 304개의 모든 개체에 대해서 수행하고, k가 5와 7일 때도 같은 방법으로 수행한다. 표 4.2로부터 수령자 파일의 1번째, 218번째, 304번째 개체에 대해 k가 1, 3, 5, 7일 때 INDUS에 매칭 되는 값을 정리한 결과 표 4.4와 같다. INDUS와 같이 연속형 유일변수인 AGE도 위와 같은 방법으로 매칭을 수행한다.

(2) 범주형 변수의 통합

제공자 파일의 유일 변수 중 범주형인 CHAS의 매칭의 경우, 매칭 하고자 하는 CHAS를 반응변수로 하여 로지스틱 회귀모형을 적합시킨다. 이 추정된 회귀모형을 이용해 제공자 파일뿐만 아니라 수령자 파일에서 구하고자 하는 확률의 예측값을 구한 후, 연속형 반응변수에서와 같이 수령자 파일의 하나의 개체에 대해서 202개의 제공자 파일의 모든 개체에 대해 예측확률의 차이를 구한다. 그리고 각 수령자의 개체에 대해 예측치의 차이가 상대적으로 작은 제공자의 CHAS 값을 이용해 매칭이 이루어진다. 표 4.3는 수령자 파일의 각 개체에 대해 예측확률의 차이가 작은 것 순으로 7개의 제공자 파일 개체를 보여준다.

표4.3에서 수령자 파일의 218번째 개체는 제공자 파일의 146번째 개체와 가장 가까운 예측확률을 갖는다. 기존 회귀분석 접근법에 의하면 제공자 파일의 146번째 개체의 CHAS 값 1을 수령자 파일에 추가시킨다. 그러나 제안된 방법에 의해 가장 가까운 3개의 개체를 고려한다면 제공자 파일의 146번째, 63번째, 64번째 개체에 해당하는 CHAS값의 최빈값 $Mode\{1,0,1\} = 1$을 수령자 파일에 추가시킨다. 이런 과정을 수령자 파일의 304개의 모든 개체에 대해서 수행하고, k가 5와 7일 때도 같은 방법으로 수행한다.

표 4.3: 가장 가까운 7개의 예측치의 차이

R	1			···	218			···	304		
k	D	차이	CHAS		D	차이	CHAS		D	차이	CHAS
1	103	0.0027	0		146	0.0181	1		91	0.0014	0
2	181	0.0037	0		63	0.0265	0		69	0.0023	0
3	172	0.0068	0		64	0.0575	1		53	0.0065	0
4	142	0.0091	1		144	0.0996	0		171	0.0081	0
5	150	0.0145	0		149	0.1230	1		94	0.0098	0
6	92	0.0163	0		60	0.1301	0		172	0.0111	0
7	91	0.0164	0		147	0.1346	0		174	0.0132	0

R:수령자파일의 개체, D:제공자파일의 개체, CHAS:제공자파일의 실제값을 나타냄.
차이 $= |\hat{P}_R - \hat{P}_D|$

표 4.3부터 수령자 파일의 1번째, 218번째, 304번째 개체에 대해 k가 1, 3, 5, 7일 때 CHAS에 매칭 되는 값을 정리한 결과 표 4.4와 같다.

표 4.4: INDUS와 CHAS의 매칭결과 (Z'_{INDUS} , Z'_{CHAS})

k \ R	INDUS의 매칭결과(Z'_{INDUS})					CHAS의 매칭결과(Z'_{CHAS})				
	1	···	128	···	304	1	···	128	···	304
1	8.140		18.100		8.560	0		1		0
3	8.727		18.100		7.030	0		1		0
5	8.492		18.100		9.846	0		1		0
7	7.131		18.100		9.861	0		0		0

4.1.3. 정확도 평가

데이터 통합 알고리즘들이 실제값을 얼마나 잘 추정하는지 평가하기 위해 본 연구에서는 통합된 변수의 실제값과 통합 알고리즘을 통하여 추가된 값을 비교하였다. 정확도의 측도로 연속형 변수에 대해서는 평균제곱오차(MSE)를 범주형 변수에 대해서는 오분류율(error rate)를 사용하였다.

데이터의 충실도를 데이터의 정확도로 평가 할 수 있으므로, 이 정확성 측도가 작게 도출되는 데이터 통합 알고리즘일수록 더 효율적인 알고리즘이다.

4.2. 결과 비교

4.1절에서 설명한 방법으로 Housing 데이터에 대해 140회, Abalone, CMC, Letter Recognition 데이터에 대해 각각 20회의 반복실험을 통해 정확성 측도의 평균을 이용하여 제안된

데이터 통합 알고리즘과 기존의 알고리즘 성능을 비교하였다.

(1) Housing 데이터

표 4.5은 Housing 데이터에 대해 반복 실험을 통해 연속형 변수의 정확도(MSE)와 범주형 변수의 정확도(오분류율)를 도출한 결과이다.

연속형 변수인 INDUS의 경우 k가 1에서 7까지 증가하면서 MSE가 점차 감소한다. AGE의 경우도 MSE는 k가 증가할수록 점차 감소한다. 두 연속형 변수에서 k가 1에서 3으로 증가할 때 MSE의 감소량이 다른 구간에 비해 상당히 크다는 것을 확인할 수 있다. 범주형 변수인 CHAS의 경우 k가 1에서 3으로 증가할 때 오분류율이 증가하고 있으므로 연속형 변수들과는 달리 k의 증가에 따른 규칙적인 패턴이 발견되지 않아 k의 영향을 단정짓기 힘들다.

표 4.5: Housing 데이터의 비교 실험 결과

변수형태	연속형								범주형			
변수명	INDUS				AGE				CHAS			
k / 실험	1	3	5	7	1	3	5	7	1	3	5	7
1	16.4	11.1	10.7	10.9	487.8	357.6	315.4	307.0	6.58	8.22	6.91	7.24
2	20.3	12.7	11.4	11.6	420.9	290.1	262.5	250.8	7.57	12.17	8.22	8.22
3	23.0	15.0	12.6	12.3	507.8	307.4	263.7	248.0	8.22	12.83	8.88	8.88
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
138	22.1	13.7	12.5	12.2	461.6	304.0	281.5	278.8	7.24	9.87	7.57	8.22
139	21.2	13.4	12.5	11.7	488.4	306.4	281.0	258.6	7.57	12.17	7.24	8.88
140	20.1	14.5	14.1	13.7	399.6	299.5	266.8	244.8	9.21	14.14	6.91	6.58
MSE	**22.6**	**15.2**	**13.7**	**13.1**	**443.1**	**303.7**	**275.2**	**263.6**	**7.25**	**12.74**	**8.32**	**7.49**
차이	7.4	1.5	0.5		139.3	28.6	11.6		-5.49	4.41	0.83	

연속형 변수: MSE, 범주형 변수: 오분류율 구함.

(2) Abalone, Contraceptive Method Choice(CMC), Letter Recognition 데이터

Abalone, CMC, Letter Recognition 데이터를 이용한 통합과정의 반복실험을 통한 k에 따른 데이터 통합의 정확도를 평가한 결과는 표 4.6과 같다. 연속형 변수의 경우 MSE값의 변화를 통해 k를 1개 사용할 때보다 상대적으로 유사한 여러 개체를 고려하여 데이터 통합을 수행할 때, 더 정확한 데이터 통합이 이루어지고 있음을 알 수 있다. 그리고 범주형 변수인 CMC 데이터의 Media변수의 경우 앞의 Housing 데이터의 범주형 변수에서 보여준 결과와는 달리 k가 클수록 정확한 데이터 통합을 수행하는 결과를 보여준다. 이렇게 두 데이터의 범주형 변수에서 다른 결과가 도출되는 이유를 살펴본 결과, 데이터 파티션의 변수 분리시 범주형 변수에 영향력 있는 공통변수가 선택되는지 여부에 달려있다고 판단된다.

앞의 Housing 데이터의 범주형 변수 CHAS는 공통변수 선택에 있어 아무런 영향을 끼치지 않고 있으므로 데이터 통합과정에서 회귀모형에 포함되는 공통변수의 수가 적거나 유의하지 않아 절편이 없는 모형을 설정한 경우가 대부분이었으나, CMC 데이터의 범주형 변수 Media는 5개의 공통변수중 2개를 결정짓는 역할을 하고 있으므로 데이터 통합과정에서 의미 있는 회귀모형을 적합할 수 있었다. 결국 신뢰할만한 회귀모형으로 도출된 예측확률로 데이터 통합을 수행한 CMC 데이터의 결론에 따라, 범주형 변수의 통합에서도 통합과정 중에 도출되는 회귀모형만 의미 있다면 k가 1보다 클 때 데이터 통합의 정확도가 높아질 것이라 판단된다.

표 4.6: Abalone 데이터의 비교 실험 결과

데이터명	변수형태	k / 변수명	MSE (차이)			
			1	3	5	7
Abalone	연속형	Length	0.000636 (0.000204)	0.000431 (0.000038)	0.000394 (0.000014)	0.000379
CMC	연속형	H-edu	0.819 (0.293)	0.526 (0.056)	0.469 (0.022)	0.448
		Nchild	7.126 (2.421)	4.705 (0.376)	4.330 (0.192)	4.138
	범주형	Media	10.49 (2.06)	8.43 (0.32)	8.11 (0.17)	7.94
Letter	연속형	onpix	3.43 (0.94)	2.49 (0.19)	2.30 (0.08)	2.22
		x2ybr	6.67 (1.89)	4.78 (0.41)	4.37 (0.16)	4.22
		xy2br	5.83 (1.54)	4.30 (0.31)	3.98 (0.13)	3.85
		yegvx	4.03 (1.31)	2.72 (0.25)	2.48 (0.06)	2.42

5. 결론 및 향후 연구 방향

본 연구에서는 마이닝에 사용될 데이터의 질을 향상시키기 위한 데이터보강을 위한 방법으로 데이터 통합 기법을 살펴보고, 그 중 하나인 회귀분석접근법을 보완하기 위한 방법을 제시하였다. 기존 회귀분석접근법을 사용하는 데이터 통합과정에서는 예측치가 가장 유사한 한 개의 개체만을 사용하여 통합이 이루어지고, 이는 상대적으로 유사한 다른 개체들이 무시되어 데이터 통합의 정확성이 떨어질 수 있는 문제점을 가지고 있다. 이를 보완하기 위해 k-최근접이웃방법의 아이디어를 고려하여 가장 유사한 한 개의 개체보다는 상대적으로 유사한 여러 개체를 사용하여 데이터 통합을 수행하였다.

여러 데이터를 이용해 실험한 결과 연속형 변수를 매칭 시키고자 할 때는 k가 1일 때보다는 k가 3, 5, 7로 증가할수록 보다 정확한 데이터 통합 작업을 수행 가능함을 알 수 있었

다. 이때, 일반적으로 k가 1에서 3으로 증가 할 때 가장 큰 MSE의 감소를 보였으며 이는 어느 정도 오류를 감수하더라도 계산량을 줄이기 원한다면 데이터 통합시 고려할 개체 수를 3으로 정하는 것이 적절하다고 생각된다. 그러나, 본 연구의 실험 데이터만으로는 범주형 데이터에 대한 안정적인 결론을 얻기 힘들었으며, 여러 실험을 통해 데이터가 200보다 작은 경우에는 매 실험마다 결과의 편차가 매우 커서 신뢰할 수 없었다. 그러므로 데이터 통합 작업은 대용량의 데이터를 가지고 수행할 때 더욱 그 성능을 발휘할 것이라고 판단된다.

향후 연구 과제로는 범주형 변수의 통합, 특히 실제 데이터는 3개 이상의 범주를 가지는 변수가 많이 포함되므로 이에 관한 연구가 더 이루어져야 할 것이며, 본 논문에서 살펴본 k-최근접이웃기법과 회귀분석접근법이외에 데이터 통합에 응용할 수 있는 다른 데이터 마이닝 기법에 관한 연구도 가치 있을 것이다.

참고문헌

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.

[2004년 4월 접수, 2004년 8월 채택]