The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 509-520

A Study on a Statistical Matching Method Using Clustering for Data Enrichment[1)]

Soon Y. Kim[2)], Ki H. Lee[3)], and Sung S. Chung[4)]

Abstract

Data fusion is defined as the process of combining data and information from different sources for the effectiveness of the usage of useful information contents. In this paper, we propose a data fusion algorithm using k-means clustering method for data enrichment to improve data quality in knowledge discovery in database(KDD) process. An empirical study was conducted to compare the proposed data fusion technique with the existing techniques and shows that the newly proposed clustering data fusion technique has low MSE in continuous fusion variables.

Keywords : Clustering, Data enrichment, Data fusion Data Mining, k-Nearest Neighbor, Statistical matching.

1. 서론

대용량의 데이터로부터 의미 있는 지식을 찾아내는 것이 목적인 데이터마이닝에서 데이터의 질은 무엇보다도 중요하다. 그러나 일반적으로 데이터로부터 정보를 얻고자 할 때, 우리가 관심을 갖는 모든 변수를 모두 포함하는 질 좋은 데이터를 얻기가 어려운 것이 현실이다. 이에 대한 해결 방안으로 여러 경로와 원천으로부터 수집된 데이터로부터 데이터를 보강(data enrichment)하여 사용하는 방법이 제안되어 왔고 이는 자료를 재조사하는 방법보다 비용과 시간을 절약할 수 있을 뿐만 아니라, 때로는 더욱 신뢰성이 높은 방법이 될 수 있으며, 조사 응답자의 부담을 줄여줌과 동시에 응답자로부터 성의 있는 응답을 기대할 수 있다(U.S. Department of Commerce, 1980).

본 논문에서는 데이터 보강을 위해서 최근 유럽지역에서 관심을 가지고 연구되는 분야 중의 하나인 데이터 통합(data fusion) 기법에 관하여 고찰하고자 한다. 데이터 통합은 그 동안 데이터마이닝 분야보다 다른 영역들에서 활발하게 응용되어 왔다. 즉, 1970년에서 1990년대까지는 미국과 독일의 경제 통계학의 많은 응용분야에서, 그 후에는 유럽과 호주의 미디어 리서치 분야에서 데이

1) 본 연구는 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음
2) 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보과학과 박사과정
E-mail : rabbit@chonbuk.ac.kr
3) 전라북도 전주시 완산구 효자동 3가 1200, 전주대학교 경영학부 교수
4) 전라북도 전주시 덕진구 덕진동 1가 644-14, 전북대학교 통계정보과학과 교수

터 통합은 매우 인기 있고 이슈가 되는 주제였다. 이처럼, 과거 30여 년 동안 데이터 통합에 대해 꾸준한 연구가 있어 왔으나 이는 근본적으로 KDD 프로세스에 다양하게 적용되지는 않았다(van der Putten et al., 2002).

데이터 통합은 자료로부터 정보를 얻기 위한 준비단계로서 분석에 필요로 하는 변수를 추가하기 위해 사용될 수 있는 기술 중 하나이다. 이것은 보유하고 있는 데이터 파일에 필요한 변수가 없거나, 결측값이 존재할 경우 다른 원천으로부터 모아지는 데이터와 정보(information)를 통합시키는 것이라고 정의된다(Saporta, 2002).

또한 데이터 통합은 대용량 데이터의 병합(macro data set merging), 통계적 레코드 연결(statistical record linkage), 다-원천 대체(multi-source imputation) 또는 통계적 매칭(statistical matching)으로도 알려져 있다. 오늘날까지 조사(survey)에서 응답자 수와 설문 문항 수를 줄이는데 주로 이용되고 있는데, 예를 들면 Belgian National Readership의 상품과 미디어에 관한 조사에서 각 10,000명을 포함하는 다른 두 그룹의 응답자를 조사하여 이를 하나의 자료로 통합하는데 사용하기도 하였다. 이와 같은 과정에서 각각의 응답자에게 들이는 시간과 비용을 줄일 수 있었고, 또한 추가되는 변수는 일반적으로 예측의 질을 향상시킨 것으로 알려져 있다(van der Putten et al., 2002).

데이터 통합 기법에 관한 기존 연구들은 거리(distance)와 같은 유사성(similarity) 측도를 이용하여 가장 유사한 개체(Nearest Neighbor)를 찾거나, 회귀분석과 같은 기법을 적용한 데이터 통합 기법이 제안되었다. 최근에는 정성석 등(2004)이 회귀분석기법에 k-Nearest Neighbor(k-NN)기법을 적용하여 상대적으로 유사한 개체에 대한 정보 손실을 줄이는 방법을 제안하였다.

그러나 데이터마이닝에서 다루는 자료들이 대부분 대용량임을 감안한다면 모든 개체간의 거리를 계산하는 데이터 통합 과정은 데이터의 양(개체의 수, 통합변수의 수)이 늘어날수록 계산량이 기하급수적으로 증가하는 부담을 갖고 있다.

본 연구에서는 대용량의 데이터를 분석하는 작업과정에서 계산량과 수행시간이 증가하는 단점을 보완하고 데이터 통합의 효율성을 높이고자 통합과정에 사용되는 개체들 중에서 유사한 것들을 몇 개의 집단으로 그룹화 하여, 각 그룹별로 통합과정을 수행하는 군집화(clustering) 통합 알고리즘을 제안하고자 한다.

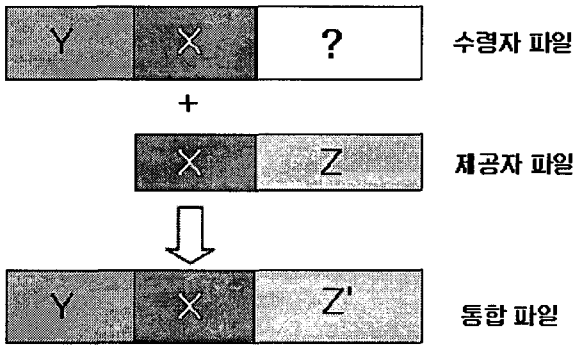
제안된 방법은 데이터의 군집화를 수행한 후 각 군집 내에서 회귀분석기법과 k-NN기법을 적용함으로써 통합에 사용되는 계산량과 수행시간의 절약을 가져올 뿐만 아니라 상대적으로 유사한 개체에 대한 정보 손실을 줄임으로써 기존의 통합 알고리즘에 비해 데이터 통합의 효율성을 높이는 방법이다. 제안하는 군집화 통합 알고리즘과 기존의 알고리즘의 효율성을 비교하기 위해 실제 데이터에 데이터 통합 기법을 적용하여 본 결과 제안하는 알고리즘이 보다 정확한 작업을 수행함을 알 수 있었다.

본 논문의 2절에서는 데이터 통합을 위한 데이터의 구조와 기존의 데이터 통합기법을 기술하였으며, 3절에서는 본 연구에서 제안하는 군집화 통합 기법을 제안하였다. 4절에서는 제안한 방법을 실제 데이터에 적용하고 그 효율성을 기존의 방법과 비교하였고 마지막으로 5절에서는 결론 및 향후 연구방향에 대해 논의하였다.

2. 데이터 통합의 구조와 데이터 통합 기법

2.1 데이터 통합의 구조

데이터 통합의 구조를 살펴보면 다음 <그림 2-1>과 같다.



<그림 2-1> 데이터 통합 개념도

통합과정에 의해 확장될 데이터 파일을 수령자 파일(recipient file), 통합을 위해 사용될 추가 정보와 같은 데이터 파일을 제공자 파일(donor file)이라 한다. 두 데이터 파일에 공통적으로 포함하는 변수를 공통변수(common variable)라 하며 X로 표시한다. 그리고 각 파일에 유일하게 존재하는 변수를 유일변수(unique variable)라 하며 수령자 파일에서는 Y, 제공자 파일에서는 Z로 표시한다.

데이터 통합은 공통변수 X를 이용하여 제공자 파일의 유일변수 Z를 수령자 파일에 추가하여 통합된 파일(fused file)을 완성하는 작업을 의미한다. 즉, 공통변수 X를 이용하여 수령자 파일의 각 개체에 가장 가까운 제공자 파일의 개체를 선택하여, 선택된 개체의 유일변수 Z를 수령자 파일에 추가한다. 이때 추가된 변수를 Z'로 표시하며 이를 통합변수 (fusion variable)라고 한다.

2.2 데이터 통합 기법

일반적인 데이터 통합과정은 수령자 파일의 한 개체와 제공자 파일의 모든 개체 사이의 거리를 계산한 후, 그 중 가장 거리가 가까운 제공자 파일의 개체를 선택하여 수령자 파일에 추가시키는 것이다.

데이터 통합 알고리즘이 유용한 결과를 도출하기 위해서는 제공자 파일은 수령자 파일을 대표할 수 있어야 하나 반드시 두 데이터가 같은 모집단에서 추출될 필요는 없고, 공통변수 X가 주어졌을 때, 유일변수인 Y와 Z사이에 조건부 독립관계가 성립되어야 한다(van der Putten et al., 2002). 조건부독립성은 데이터 통합 과정에서 유용한 가정으로 Rässler(2002)가 제시한 회귀분석접근법(regression approach)을 활용하여 판단할 수 있다. 본 절에서는 기존 연구의 통계적 매칭 알고리즘인 k-NN방법, 회귀분석방법과 정성석 등(2004)이 제안한 회귀분석과 k-NN방법을 혼합한 데이터 통합기법을 살펴보겠다.

2.2.1 k-최근접이웃(k-NN) 알고리즘

최근접이웃방법은 수령자 파일의 한 개체에 가장 가까운 제공자 파일의 한 개체를 이용하여 통합에 사용하는 방법이고, 상대적으로 유사한 제공자 파일의 k개의 개체를 이용하여 통합에 사용하는 방법이 k-최근접이웃방법이다. 이는 van der Putten et al.(2002)에 의해 제시된 기법으로 공통변수 X를 이용하여 수령자 파일의 각 개체에 대해 제공자 파일의 모든 개체와의 거리를 계산한다. 이중 가장 가까운 제공자 파일의 k개의 개체를 선택한 후, 선택된 k개 개체에 해당하는 제공자 파일의 유일변수 Z를 이용하여 수령자 파일의 각 개체에 통합 변수를 추가시킨다. 이때 유일변수가 연속형인 경우 k개 Z값의 평균(mean)을, 범주형이면 k개 Z값의 최빈값(mode)을 이용한다.

2.2.2 회귀분석 알고리즘

회귀분석을 이용하는 방법은 제공자 파일에서 회귀모형을 추정한 후, 추정된 회귀모형을 이용하여 수령자 파일과 제공자 파일에서 예측치를 구한다. 그리고 두 파일의 예측치 사이의 거리가 가장 짧은 개체를 이용하여 통합에 사용하는 방법이다. 이는 Ingram et al.(2000)에 의해 제시된 기법으로 제공자 파일의 유일변수 Z를 목표변수로, 제공자 파일의 공통변수 X를 설명변수로 하여 회귀모형을 추정한다. 추정된 회귀모형을 수령자 파일과 제공자 파일에 적용하여 각 파일에서 유일변수 Z의 예측값 $\hat{Z}$을 구한 후, 이 예측값 $\hat{Z}$을 이용하여 수령자 파일의 각 개체와 제공자 파일의 모든 개체사이의 거리를 계산한다. 그리고 예측값 사이의 거리가 가장 가까운 제공자 파일의 개체 하나를 선택하고, 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 이용하므로 데이터 통합방법에서 회귀분석과 같은 예측평균매칭(predicted mean matching)기법은 좋은 성능을 나타낸다(Ingram et al., 2000).

2.2.3 수정된 데이터 통합 알고리즘

정성석 등(2004)이 제안한 이 데이터 통합기법은 상대적으로 유사한 개체에 대한 정보 손실을 줄여 데이터 통합의 성능을 높이고자 회귀분석기법에 k-NN기법을 결합하여 통합변수를 추가시키는 방법이다.

제공자 파일의 유일변수 Z를 목표변수로, 제공자 파일의 공통변수 X를 설명변수로 하여 회귀모형을 추정한 후, 추정된 회귀모형을 수령자 파일과 제공자 파일에 적용하여 각 파일에서 유일변수 Z의 예측값 $\hat{Z}$을 구한다. 두 파일의 예측값을 이용하여 수령자 파일의 각 개체에 대해 모든 제공자 파일 개체와의 거리를 구한다. 수령자 파일의 각 개체에 예측값 사이에 거리가 가장 가까운 k개의 제공자 파일의 개체를 선택하고, 선택된 개체의 유일변수 Z를 수령자 파일의 개체에 추가한다. 이때 유일변수가 연속형인 경우 k개 Z값의 평균(mean)을, 범주형이면 k개 Z값의

최빈값(mode)을 이용한다.

3. 군집화 데이터 통합기법

본 연구에서 제시한 군집화(Clustering) 방법을 이용한 데이터 통합기법은 데이터 통합과정을 수행하기 전에 우선 제공자 파일에 비계층적 군집화 방법의 하나인 c-평균 군집화(일반적으로는 k-means clustering이라 표현하지만 본 논문에서는 앞의 k-NN방법의 k와 혼동을 피하기 위해 k 대신 c를 사용한다)를 적용하여 유사한 개체끼리 몇 개의 집단으로 그룹화하고, 수령자 파일을 가장 가까운 그룹에 할당한 다음 각 그룹별로 통합과정을 수행하는 방법이다. 이 방법을 자세히 살펴보면 다음과 같으며 군집화 데이터 통합의 구조는 <그림 3-1>과 같다.

step 0. 제공자 파일을 이용하여 주성분 분석을 실시하여 군집의 수(c)를 결정한다.

step 1. step 0 에서 제공된 군집수를 이용하여 제공자 파일에서 공통변수 X를 이용하여 c-평균 군집화(c-means clustering)과정을 수행한다.

step 2. 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.

step 3. i ($i=1,\cdots,c$)번째 군집에서 제공자 파일의 유일변수 Z중 임의의 s번째 변수를 목표변수로 공통변수 X를 설명변수로 하여 회귀모형을 추정한다.

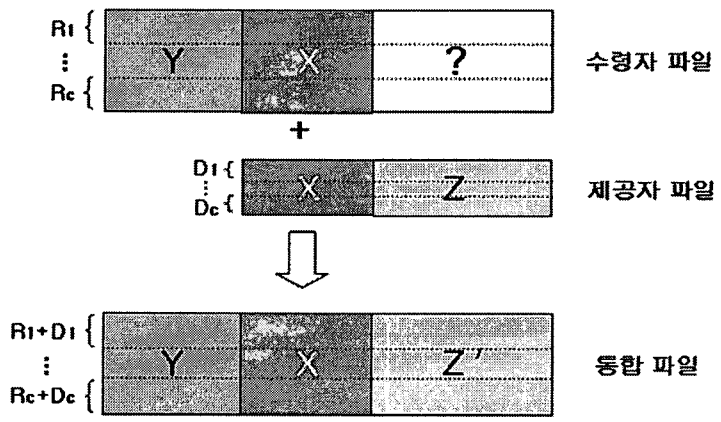
step 4. i ($i=1,\cdots,c$)번째 군집에서 추정된 회귀모형을 해당 군집의 수령자 파일과 제공자 파일에 적용하여 각 파일에서 s번째 유일변수 Z의 예측값 $\hat{Z}$을 구한다.

step 5. i ($i=1,\cdots,c$)번째 군집에서 두 파일에서의 s번째 유일변수 Z의 예측값을 이용하여 수령자 파일의 각 개체에 대해 해당 군집의 모든 제공자 파일의 개체와의 거리를 구한다.

step 6. step 5에서 구한 예측값 사이의 거리를 이용하여 수령자 파일의 각 개체에 가장 가까운 제공자 파일에 해당하는 k개의 개체를 선택한다.

step 7. i ($i=1,\cdots,c$)번째 군집에서 선택된 제공자 파일의 k개의 개체들의 s번째 유일변수 Z들의 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.

step 8. s ($s=1,2,\cdots,S$)번째 유일변수에 대해 step 1에서 step 7까지를 반복적으로 적용하여 하나의 통합된 데이터 파일을 형성한다.



<그림 3-1> 군집화 데이터 통합 개념도

4. 실제 자료에의 적용

3절에서 제안된 방법과 기존의 방법의 데이터 통합 성능을 비교하기 위해 실제 데이터를 이용하여 다음과 같은 사례연구를 하였다.

사례연구에 사용된 데이터는 UCI Repository(Blake and Merz, 1998)에 있는 Abalone, Letter Recognition 그리고 Pen-Based Recognition of Handwritten Digit 데이터로서 목표변수를 제외하면 모두 연속형 변수이다. 이를 이용하여 정성석 등(2004)이 제안한 데이터 통합 알고리즘과 본 연구에서 제안하는 군집화 데이터 통합 알고리즘의 성능을 k가 1일 때, 그리고 3, 5, 7로 증가함에 따라 정확성 측도의 값을 비교하여 새롭게 제안한 데이터 통합 알고리즘의 효율성을 비교하고자 하였다.

4.1 데이터 실험 과정

본 연구에서 데이터 통합 기법의 효율성 비교는 통합변수의 실제값과 통합된 값의 차이인 평균제곱오차(MSE)를 이용하게 된다. 모의실험을 위해 하나의 데이터를 파티션하여 수령자 파일과 제공자 파일로 분리하였다.

첫 번째 단계인 데이터 파티션은 수령자 파일과 제공자 파일이 포함할 변수의 분리와 개체의 분리 두 가지를 포함한다. 수령자 파일과 제공자 파일에 포함할 데이터의 비율은 Yoshizoe and Araki (1999)에서와 같이 60%대 40%로 하고, 데이터의 분리는 단순임의(simple random) 방법을 사용하였다. 그리고 각 파일이 포함할 변수를 분리하는 과정은 통계적 매칭에서 가장 중요하게 여겨지는 조건부독립성을 근거로 하였다. Rässler(2002)가 제시한 조건부독립성을 회귀분석접근법(regression approach)으로 판단하는 방법을 역으로 이용하여 조건부독립성 가정이 유지되도록 변수를 분리하였다

각 파일이 포함할 변수를 분리하기 전에 최종 분석의 목표변수(target variable)는 데이터 통합에 영향을 주거나 받지 않도록 하기 위해 수령자 파일의 유일변수 Y에 포함하였다. 즉, 다음과

같은 회귀모형 $Z=\beta_0+\beta_{ZX,Y}X+\beta_{ZY,X}Y+\epsilon$에서 $\beta_{ZY,X}=0$이면 $\rho_{ZY|X}=0$과 같은 관계로부터 유의수준 0.05로 목표변수가 설명변수로 유의하지 않은 반응변수를 제공자 파일의 유일변수 Z로 선택하였다. 그리고 공통변수 X는 제공자 파일의 유일변수로 선택된 변수를 반응변수로 하여 유의한 설명변수를 선택하였다. 그리고 제공자 파일의 유일변수에도 포함되지 않고 공통변수에도 포함되지 않는 변수는 수령자 파일의 유일변수 Y에 포함시켰다.

각 데이터의 파티션 결과는 다음 <표 4-1>과 같다.

<표 4-1> 실험 데이터의 파티션 결과

데이터	변수		개체수			제공자파일 유일변수(Z)	공통변수(X)	수령자파일 유일변수(Y)	군집수
	연속형	범주형	전체	수령자 파일	제공자 파일				
Letter Recognition	16	1	20,000	12,000	8,000	x_box, y_box, width, high, onpix, x2ybr, 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
Abalone	7	1	4,177	2,506	1,671	Length	Diameter Sweight Vweight	Sex, Height, Wweight, **Rings****	2

* : Pen-Based Recognition of Handwritten Digit.

** : 각 데이터의 목표변수(target variable)를 의미함

두 번째 단계인 통합과정에서 본 연구에서 제안하는 군집화 통합 알고리즘을 사용하기 위해서는 파티션된 데이터 중 제공자 파일을 유사한 개체끼리 군집화를 형성한다. 군집화는 c-평균 군집화를 이용하고, 이때 군집의 수는 다변량 통계분석 기법중 주성분 분석(principle component analysis)을 이용하여 고유값의 크기가 1이상인 주성분의 개수를 군집의 수로 결정하여 사용한다. 또한 두 개체사이의 유사성의 척도는 연속형 변수의 경우 유클리디언 거리(Euclidean distance)를 사용하였다.

Letter Recognition 데이터의 경우 4개, Handwritten Digits 데이터의 경우 3개의 군집을 이용하여 데이터 통합을 수행하였다. 그러나 Abalone 데이터의 경우 고유값이 1이상인 개수가 1개였다. 즉, 군집화가 필요 없다는 의미이나 본 연구에서는 이러한 경우에도 군집화 통합 알고리즘과 기존의 통합 알고리즘의 성능을 비교해보고자 2개의 군집을 사용하여 Abalone 데이터를 실험에 적용시켰다.

Letter Recognition 데이터를 이용하여 데이터 통합을 살펴보면 다음과 같다.

step 1. 주성분 분석에 의해 결정되어진 군집의 수를 사용하여 제공자 파일을 유사한 개체끼리 4개의 군집으로 나눈다.

step 2. 제공자 파일의 각 군집의 중심과 수령자 파일의 모든 개체사이의 거리를 계산하여 수령자 파일의 개체를 거리가 가장 가까운 군집에 할당한다.

step 3. 각 군집별로 다음과 같은 통합과정을 수행한다. 제공자 파일의 유일변수인 x2ybr변수의 통합의 경우, 통합될 변수가 연속형 이므로 통합과정에서 이상치의 영향을 배제하여 회귀모형으로부터 구해진 두 파일의 예측치 차이가 2이하인 개체만을 통합에 고려하기 위해 회귀분석 전에 x2ybr를 S_x2ybr로 표준화한 후 S_x2ybr를 반응변수, 공통변수 X를 설명변수로 하여 제공자 파일에서 회귀모형을 적합시킨다. 또한 통합에 사용될 회귀모형에 설명력 있는 공통변수만 포함되도록 단계적(stepwise) 변수선택을 수행하였다.

step 4. 적합된 회귀식을 이용하여 제공자 파일과 수령자 파일에서 S_x2ybr의 예측값을 구한다.

step 5. 수령자 파일의 하나의 개체에 대해서 제공자 파일의 모든 개체에 대해 예측값의 차이(거리)를 구한다.

step 6. 그리고 각 수령자 파일의 개체에 대해 예측값의 차이가 상대적으로 적은 제공자 파일의 k개의 개체를 선택한다($k=1,3,5,7$).

step 7. k개 예측값의 평균을 이용하여 데이터 통합을 수행한다.

본 연구에서는 차이가 가장 작은 1, 3, 5, 7개의 제공자 파일의 개체를 각각 사용하여 각 군집별로 위에서 언급한 통합과정을 수행하였고 각 군집을 결합하여 최종 통합된 파일을 만들었다.

세 번째 단계인 정확도 평가에서는 데이터 통합 알고리즘들이 실제값을 얼마나 잘 추정해내는지 평가하기 위해 통합된 변수의 실제값과 통합 알고리즘을 통하여 추가된 값을 비교해 보았다. 정확도의 측도로 연속형 변수에 대해서 평균제곱오차(MSE)를 사용하였다. MSE가 작은 통합 알고리즘일수록 더 효율적인 알고리즘이다.

4.2 통합 결과 비교

4.1절에서 설명한 방법으로 Letter Recognition, Abalone 그리고 Handwritten Digits 데이터에 대해 데이터를 파티션한 후 데이터 통합을 수행하였다. 파티션된 데이터가 달라짐에 따라 실험 데이터와 실험 결과가 달라진다. 그러므로 정성석 등(2004)에서 사용한 반복횟수 20회를 사용하여 실험을 20회 반복적으로 수행하여 구한 MSE들의 평균을 기준으로, 제안된 군집화 데이터 통합 알고리즘과 정성석 등(2004)의 알고리즘의 효율성을 비교하였다.

통합과정의 반복실험을 통한 데이터 통합의 정확도를 평가한 결과는 k값에 따라 두 방법의 각 변수의 MSE를 비교한 <표 4-2>에 주어져 있고 이들 중 몇 개의 변수에 대한 MSE를 <그림 4-1>, <그림 4-2> 그리고 <그림 4-3>에 표시하였다.

통합에 사용하는 개체의 수의 증가에 따른 결과는 두 방법 모두 k가 1에서 7까지 증가할수록 MSE가 점차 감소한다. k가 1에서 3으로 증가할 때 MSE의 감소량은 다른 구간에 비해 상당히 크다는 것을 확인할 수 있다. 즉, k를 1개 사용할 때보다 상대적으로 유사한 여러 개체를 고려하여 데이터 통합을 수행할 때, 더 정확한 데이터 통합이 이루어지고 있음을 알 수 있다.

군집화 통합 알고리즘과 기존의 통합 알고리즘에 의한 결과를 살펴보면 다음과 같다.

Letter Recognition 데이터의 경우 onpix와 high변수를 제외하면 군집화 데이터 통합 알고리즘의 MSE가 더 작음을 알 수 있다. onpix와 high 변수의 경우는 k가 5 또는 7로 증가한 경우 기

존의 알고리즘보다 군집화 데이터 통합 알고리즘의 MSE가 더 작음을 알 수 있었으나 거의 비슷한 정확도를 나타냄을 알 수 있다.

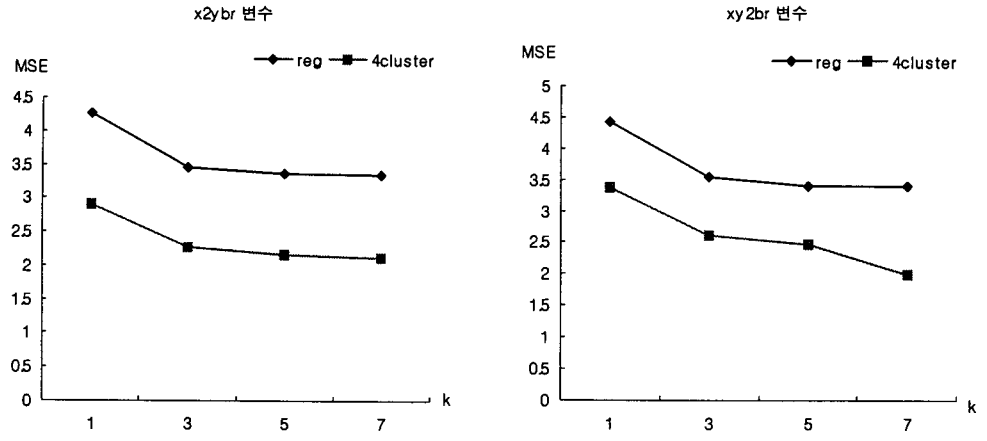
Handwritten Digit 데이터의 경우 모든 변수에 대하여 제안한 군집화 데이터 통합 알고리즘의 MSE가 월등히 작아 군집화 통합 알고리즘이 더 정확한 데이터 통합을 수행함을 알 수 있다.

Abalone 데이터의 경우 군집화 데이터 통합 알고리즘이 기존의 알고리즘보다 작은 MSE를 갖으나 두 알고리즘이 거의 비슷한 통합 효율성을 갖고 있음을 알 수 있었다. 이는 Abalone의 경우 군집을 나눌 때 사용한 고유값이 1이상인 주성분의 개수가 1로서 군집화가 필요 없었으나 2개의 군집을 사용하여 통합과정을 수행하였기 때문에 통합의 효율성이 뚜렷하게 나타나지 않은 것으로 판단된다.

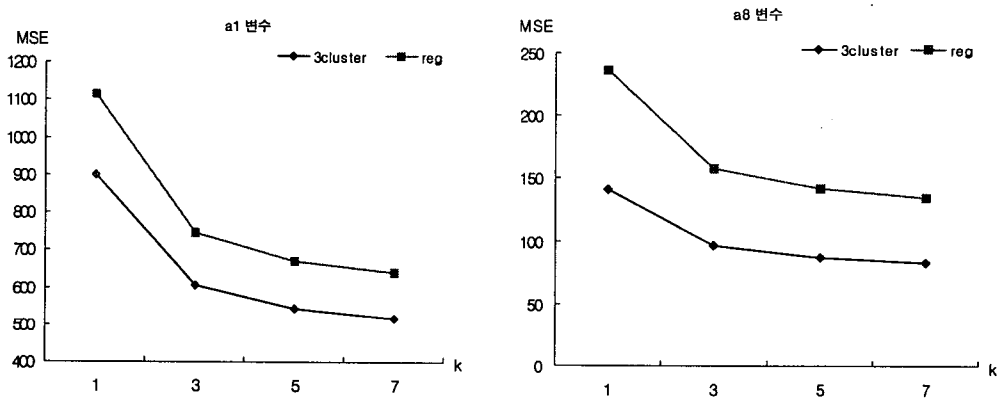
<표 4-2> 통합기법의 정확도 비교

데이터명	변수명 \ k / 통합기법		MSE			
			1	3	5	7
Letter Recognition	x2ybr	reg+knn	4.26628	3.45709	3.35354	3.33558
		Cluster	2.89559	2.25587	2.15822	2.11776
	xy2br	reg+knn	4.42429	3.53794	3.40872	3.3913
		Cluster	3.36229	2.60125	2.46880	1.99142
	yegvx	reg+knn	3.44726	2.56054	2.39379	2.32045
		Cluster	2.88942	2.13041	2.41729	1.93350
	x_box	reg+knn	3.32141	2.43658	2.27565	2.20890
		Cluster	3.03555	2.20894	2.04659	1.98200
	y_box	reg+knn	12.3752	9.35064	8.89108	8.73279
		Cluster	11.6229	8.73658	8.27886	8.12529
	width	reg+knn	3.07275	2.32746	2.20523	2.16922
		Cluster	2.99889	2.21389	2.08219	2.03548
	high	reg+knn	5.31536	4.11682	3.97210	3.94985
		Cluster	5.32382	4.05162	3.87330	3.82841
	onpix	reg+knn	2.72629	2.04614	1.94107	1.90977
		Cluster	2.81865	2.06460	1.91978	1.86180
Handwritten Digits*	a1	reg+knn	1114.874	746.078	671.648	640.587
		Cluster	898.964	603.737	543.153	517.729
	a3	reg+knn	575.477	383.154	344.225	327.571
		Cluster	508.114	339.780	306.107	291.684
	a6	reg+knn	240.229	160.185	144.021	136.913
		Cluster	165.330	111.268	100.769	95.877
	a8	reg+knn	236.354	158.098	141.800	134.876
		Cluster	141.325	96.575	87.162	83.043
	a13	reg+knn	399.573	263.901	237.589	226.635
		Cluster	281.934	191.325	173.716	165.862
Abalone	Length	reg+knn	0.00062	0.00043	0.00039	0.00038
		Cluster	0.00061	0.00042	0.00038	0.00037

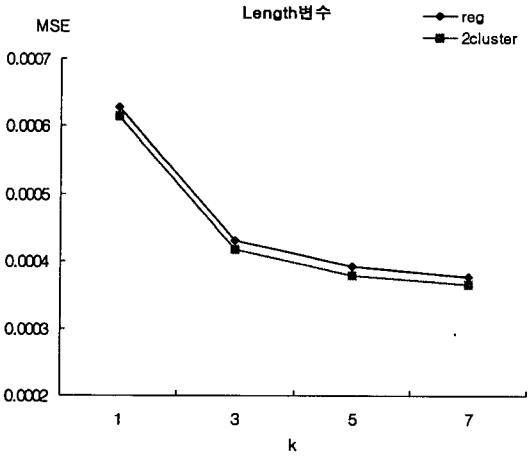
reg+knn : 정성석 등(2004)의 통합 알고리즘.
Cluster : 제안하는 군집화 통합 알고리즘.



<그림 4-1> Letter Recognition 데이터의 MSE



<그림 4-2> Handwritten Digits 데이터의 MSE



<그림 4-3> Abalone 데이터의 MSE

5. 결론 및 향후 연구과제

본 연구에서는 데이터마이닝에 사용될 데이터의 질을 향상시키기 위한 방법으로 데이터 통합 기법을 살펴보고, 그 중 하나인 정성석 등(2004)이 제안한 방법에서 데이터의 양(개체의 수, 통합 변수의 수)이 늘어남에 따라 계산량과 수행시간이 기하급수적으로 증가하는 단점을 보완하기 위해 개체들 중에서 유사한 것들을 몇 개의 집단으로 그룹화 하여, 각 그룹별로 통합과정을 수행하는 군집화를 이용한 데이터 통합 방법을 제시하였다. 또한 통합 과정에서 예측값과 가장 유사한 한 개의 개체만 사용하면 상대적으로 유사한 다른 개체들이 무시되어 데이터 통합의 정확성이 떨어질 수 있는 문제점을 고려하여 가장 유사한 한 개의 개체보다는 상대적으로 유사한 여러 개체(k)를 사용하여 데이터 통합을 수행하였다.

실험 결과 정성석 등(2004)이 제안한 방법보다 본 연구에서 제안한 군집화를 이용한 통합 방법이 더 정확한 데이터 통합 작업을 수행함을 알 수 있었으며, 두 방법 모두 k가 1일 때보다 3, 5, 7로 증가할수록 보다 정확한 데이터 통합 작업을 수행함을 알 수 있었다. 일반적으로 k가 1에서 3으로 증가할 때 가장 큰 오류율의 감소를 보였다. 또한 데이터의 개수가 1,000개 미만인 경우에 군집화를 이용한 데이터 통합 방법은 매 실험마다 결과의 편차가 커서 결과를 신뢰할 수 없었다. 그러므로 데이터마이닝의 특성상 대용량의 데이터를 이용하여 작업할 때에는 군집화를 이용한 데이터 통합 방법이 그 성능을 발휘할 거라 판단된다.

향후 연구 과제로는 범주형 변수의 통합에 관한 연구가 이루어져야 할 것이다. 정성석 등(2004)이 2개의 범주를 갖는 범주형 변수의 통합 방법에 관하여 제안하였으나, 실제 데이터는 3개 이상의 범주를 갖는 변수가 많이 포함되므로 이에 관한 연구도 향후 이루어져야 할 것이다. 마지막으로 본 논문에서 살펴본 k-NN기법, 회귀분석접근법 그리고 군집분석 이외에 데이터 통합에 응용할 수 있는 다른 통계적 기법에 관한 연구도 가치 있을 것으로 생각된다.

참고문헌

[1] 정성석, 김순영, 김현진 (2004). 데이터 보강을 위한 데이터 통합기법에 관한 연구, 「응용통계연구」, 제17권, 605-617.

[2] Blake, C. L. and Merz, C. J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

[3] 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.*

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

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

[6] 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.

[7] 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.

[8] 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.

[2005년 3월 접수, 2005년 7월 채택]