커널필터링 기법을 이용한 건강비용의 효과적인 지출에 관한 군집화 분석

정 용 규* · 최 영 진** · 차 병 헌***

목 차

요약

데이터마이닝은 방대한 데이터를 기반으로 정보를 추출하는 방법으로 많은 분야에 적용하고 있으며 특히 보건의료 데이터를 다루는 기법으로 많이 활용 되고 있다. 하지만 데이터가 다양하고 방대해짐에 따라 데이터들을 완벽하게 다룰 수 있는 알고리즘이 개발되지 못한 현황이다. 따라서 본 논문에서는 군집화 알고리즘 중의 하나인 DBSCAN 알고리즘과 EM 알고리즘의 성능을 동일한 데이터에 대하여 분석을 시도하였다. 이를 위하여 DBSACN과 EM 알고리즘에 따른 변화를 Health expenditure 실험데이터의 결과를 기반으로 분석 하였고 더욱 정확한 실험과 더욱 정확한 결과를 알아내기 위하여 Kernel Filtering을 통하여 정확한 데이터분석을 시도하였다. 본 연구에서는 알고리즘의 기술적 성능을 비교한 것을 물론이고 성능을 높이기 위한 시도를 하였다. 이를 통하여 확장한 알고리즘에 따른 성능의 변화와 실험데이터의 적용결과를 기반으로 비교하고 이를 분석하게 되었다. 특히 의료기관을 이용하는 다양한 군집으로부터 데이터 레코드를 수집하여 의료 서비스에 대한 효과적인 비용 지출을 권장할 수 있도록 실험하였다.

표제어: Health Expenditure Data, EM, DBSCAN, 군집화, 회귀분석

접수일(2015년 8월 11일), 수정일(1차: 2015년 9월 23일), 게재확정일(2015년 9월 23일)

* 을지대학교 의료IT마케팅학과, ygjung@eulji.ac.kr
** 을지대학교 의료경영학과, yuzin@eulji.ac.kr
*** 교신저자, 을지대학교 임상병리학과, jabogy@eulji.ac.kr

1. 서론

비즈니스 환경의 변화에 따라 기업은 의사결정 지원을 위한 고급정보를 필요로 하게 되었으며, 비즈니스 우위를 위해 기존의 대용량 데이터베이스의 조회방법보다 우수한 분석모델을 통한 예측 데이터를 필요로 하게 되었다. 데이터마이닝을 한마디로 요약하면 “대량의 데이터 집합으로부터 유용한 정보를 추출하는 것”으로 정의된다. 데이터마이닝은 대량의 가공하지 않은 데이터로부터 알려지지 않은 새로운 정보나 유용한 패턴과 상관관계를 추출하여 의사 결정에 이용하는 작업으로써 최근 H/W와 S/W를 비롯한 IT의 기술적 발전과 더불어 많은 연구가 이루어져 왔다. 데이터마이닝은 보험사기 색출, 이탈고객 모델링, 타겟마케팅, 교차판매, 상승판매, 상품 디스플레이, 위험관리, 웹마이닝(Web Mining), 동적 웹 페이지 구성 등 다양한 산업분야에서 연구 및 활용되고 있다.

이러한 데이터마이닝 분야에서는 데이터에 대한 통계분석이나 모델링을 통하여 정보를 추출해내기 위해서 연관성(Association), 군집화(Clustering), 결정나무(Decision Tree), 신경망(Neural Network) 등의 다양한 알고리즘들이 연구되고 있는 현황이다[6]. 그러나 실제 문제에서 이러한 기법들을 적용하는 경우, 그 결과에 영향을 미치는 요인이 다수 존재하기 때문에 모든 상황에서 완벽하게 동작할 수 있는 최적의 알고리즘을 선택하는 작업은 단순한 문제가 아니다.

특히 최근 들어 건강에 관한 관심도 많아지고 따라서 건강에 사용하는 건강비용을 중요하게 되었다. 본 논문에서는 건강비용을 효과적으로 지출하는 방법을 제시하기 위하여 공공데이터를 활용하여 분석하게 되었다. 분석기법으로는 DBSCAN, EM 알고리즘을 사용하였으며 보다 정확성과 효율을 높이기 위하여 Kernel Filtering을 적용하였다.

2. 관련 연구

2.1 선형회귀모델

선형 모델은 또 다른 간단한 표현 스타일로, 출력 형태는 그저 속성 값들을 더해놓은 꼴을 갖는다. 여기서 각 속성은 가중치들이 곱해져 있다. 선형 모델을 통해 가중치로 사용하기 좋은 값을 얻을 수 있다. 여기서 출력 데이터와 입력 데이터는 모두 수치 타입이다. 통계학자들은 수치적 수량을 예측해나가는 과정을 회귀라 부르며, 회귀 모델은 선형 모델의 또 다른 이름이다. 예측할 클래스는 없지만 인스턴스들이 각 성격에 맞는 그룹들로 분할될 경우 군집 방식을 적용할 수 있다. 군집화 알고리즘은 아마도 인스턴스들이 도출된 해당 분야에서 동작하는 메커니즘을 반영할 것이다. 여기서 메커니즘은 특정 인스턴스들끼리 다른 나머지 인스턴스들보다 더욱 비슷한 성질을 같게 한다. 군집화는 본질적으로 다양한 분류 방식과 연관 규칙들을 요구한다.

예측할 클래스는 없지만 인스턴스들이 각 성격에 맞는 그룹들로 분할될 경우 군집 방식을 적용할 수 있다. 군집화 알고리즘은 아마도 인스턴스들이 도출된 해당 분야에서 동작하는 메커니즘을 반영할 것이다. 여기서 메커니즘은 특정 인스턴스들끼리 다른 나머지 인스턴스들보다 더욱 비슷한 성질을 같게 한다. 군집화는 본질적으로 다양한 분류 방식과 연관 규칙들을 요구한다.

일반적으로 군집해석은 hierarchical clustering 기법과 nonhierarchical clustering 기법으로 구분될 수 있다. hierarchical 기법에는 single linkage, complete linkage, average linkage, median, Ward 기법 등이 있으며 nonhierarchical 기법에는 K-means, adaptive K-means, K-medoids, fuzzy clustering 기법 등이 있다. 대상 자료와 분류목적에 따라 기법을 선택해야 하므로 한 가지 이상의 기법을 적용하여 결과를 면밀하게 해석 및 비교 하는 것이 중요하며 모의실험을 통해

군집해석의 안정성을 평가할 수 있다. 군집해석에는 여러 가지 알고리즘이 있지만 single linkage기법 또는 nearest neighbor기법에 대하여 설명하면 다음과 같다. 우선 세트 내 의각개체와 다른 개체간의 거리를 계산한다. 처음에는 모두 하나의 개체로 구성된 집단들 뿐이다. 가장 가까운 짝끼리 결합을 시키게 되며 그 거리가 또 다른 개체와 같다면 세 개체 모두 하나로 병합 되며 거리가 같은 개체가 더 많을 경우에도 마찬가지 방법으로 병합된다. 다음으로 가장 작은 짝 또는 그룹들이 형성되어 간다. 이 과정을 군집해석의 다른 방법인 division과 구별하여 agglomeration이라고 부른다. 실제로 각 개체는 하나 이상의 특성을 가질 수 있으므로 거리에는 모든 특성들을 고려하여야 한다. 이 거리의 척도를 generalized Euclidean distance라고 한다. 예를 들어 특성을 정량화한 p개의 변수 Z_1, Z_2, …, Z_p가 있다면 개체 i와 j간의 거리는 다음 식과 같이 주어진다.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (z_{ik} - z_{jk})^2} \qquad (1)$$

만일 p = 2라면 그 거리는 Pythagoras 정리와동일하다. 군집해석을 수행하기 위한 한 가지 방법은 주어진 세트에 대한 주요소해석 principal component analysis)을 시작하는 것이다. 만일 두 개의 주요소(principal components)로 대부분의 분산을 설명할 수 있다면 다음 단계로 개체 들을 그룹으로 분할하는 방법을 검토할 수 있다. 그러나 이 방법은 단지 guide로 선택된 것이므로 이 방법이 자료의 직접 해석과 다를 수 있음을 주의해야 하며 주 요소가 고려하는 특성을 완전히 대표할 수는 없다.

2.2 DBSCAN

DBSCAN은 유클리드 거리를 이용해 군집 안에 존재하는 인스턴스들 중 어떤 인스턴스가 함께 있는지 결정하지만, K-means와는 달리 자동으로 군집의 개수를 결정하며, 임의의 형태를 갖는 군집을 찾고 외톨이 항을 포함한다. 적어도 점들의 최소 개수를 포함하는 것은 군집이라 정의하며, 이런 군집이 포함하는 모든 점의 쌍들은 사용자가 정의한 거리인 ϵ보다 가까운 상태에 있는 군집 안에 존재하는 일련의 점들과 연결된다.

DBSCAN(Density Based Spatial Clustering of Applications with Noise) 은 밀도를 기반으로 하는 클러스터링 알고리즘으로서, 잡음(noise)을 포함한 공간 데이터를 다루는데 적합하며 다양한 모양(arbitrary shape)과 크기(shape)의 클러스터를 구분할 수 있다. 이때 클러스터와 잡음은 직관적으로 점들의 밀도(density)를 기준으로 구분한다. 이를 위해 다음의 몇 가지를 정의한다.

(1) 한 점 p의 Eps-neighborhood 는 p로부터 반경 Eps 내에 있는 이웃의 집합이다.
즉, Eps-neighborhood of a point p, NEps(p) = {q ∈ D¦dist (p, q) ≤ Eps}이다.

(2) 한 점 p가 점 q로부터 directly density-reachable 하다는 의미는 p ∈ NEps (q): p가 q의 neighbor에 있어야 하고, [NEps (q)] ≥ MinPts (core point condition): q가 core point이다.
즉 충분한 neighbor를 갖는 것을 말한다.

(3) 한 점 p가 점 q로부터 density-reachable 하다는 의미는 p로부터 q까지의 p_{i+1}이 directly density-reachable from p_i인 chain이 존재한다는 의미이다.

2.3 EM 알고리즘

EM 알고리즘은 초기 값을 추측하는 일부터 시작해 이 변수들을 이용해 각 인스턴스에 대한 군집 확률을 계산하고, 이 확률을 이용해 매개변수 값을 다시 추정하는 식으로 반복하는 것을 말한다. 인스턴스 i가 군집 A에 속할 확률이 w_i라면 군집 A의 평균과 표준 편차는 다음과 같다.

$$\mu_A = \frac{W_1X_1 + W_2X_2 + \cdots + W_nX_n}{W_1W_2\cdots + W_n} \quad (2)$$

$$\sigma_A^2 = \frac{W_1(X_1-\mu)^2 + W_2(X_2-\mu)^2 + \cdots + W_n(X_n-\mu)^2}{W_1 + W_2 + \cdots + W_n} \quad (3)$$

데이터 집합으로부터 해당 데이터가 추출될 전반적인 우도를 계산함으로써 얼마나 근사하게 수렴하는지 알 수 있다. 이런 전반적인 우도 값은 다음과 같이 개개의 인스턴스 i와 확률 값을 곱해서 구한다.

$$\prod_i (p_a \Pr[x_i|A] + p_b \Pr[x_i|B]) \quad (4)$$

2.4 Kernel Filter

데이터를 커널 형태로 변환한다. 즉, 새로운 데이터 집합을 출력하는데, 여기에는 이전과 동일한 개수의 인스턴스가 들어있다. 여기서 각 인스턴스의 값은 기존 인스턴스 한 쌍에 대해 커널 함수가 내린 평가 결과가 된다. 기본적으로 모든 값은 평균이 0에 가깝게 변환된다. 분산은 단위 값인 1을 갖게 스케일링 되지 않더라도 다양한 필터들을 지정할 수 있다.

변환 필터의 마스크는 다양한 크기를 가질 수 있으므로 적절한 마스크의 크기를 선택하는 것이 중요하다. 콘볼루션을 거쳐 도출 되는 최대 값은 에지의 방향을 따르는 방향성 정보를 나타내고 최소 값 1은 에지 성분을 가로지르는 방향성 정보를 나타내므로 이들을 이용하여 적절한 필터링을 수행할 수 있다.

3. 실험 및 실험 결과

3.1 실험데이터

현대 사회에서 건강 비용은 매년 증가하는데, 그 이유는 의료기술의 발달과, 의료시설의 증대, 병원, 제약회사 등 건강에 관련된 회사, 제품에 대한 관심을 더 가지게 되었다. 또한 건강에 대한 관심이 증가하면서 그에 따라서 지출되는 비용을 효과적으로 사용하는데 또한 관심이 많아졌다.

실험데이터는 호주의 건강비용 지출을 정하는 속성변수로서 financial year의 Numeric 속성과 state, area of expenditure, broad source of funding, detailed source of funding의 attribute는 Nominal로 돼 있고 알고 싶은 건강지출 비용속성인 real expenditure millions는 Numeric으로 구성되어 있다. 각각의 속성에 대한 세부적인 사항은 표 1과 같다.

표 1. 실험데이터의 속성
Tab. 1. Properties of the Experimental Data

attribute	value
financial year	continuous from 2000 to 2011
state	{NSW, QLD, VIC, WA, SA, TAS, ACT, NT}
area of expentiture	{Administration, Aids and-appliances, All other medication-s, Benefit paid pharmaceuticals, Capital expenditure, Community-health, Dental services, Medical-expense tax rebate, Medical-services, Other health-practitioners, Patient transport-services, Private hospitals, Public health, Public hospitals, Research
broad source of funding	{Government, Non Government}
detailed source of funding	{Australian Government, State and local, Private health-insurance funds, Individuals, Other Non government}
real expenditure millions	continuous from 0 to 6838

Financial year는 회계연도를 의미하고 state는 호주의 지역을 의미한다. area of expentiture은 의료비용이 어디에 지출 되었는지를 알려준다.

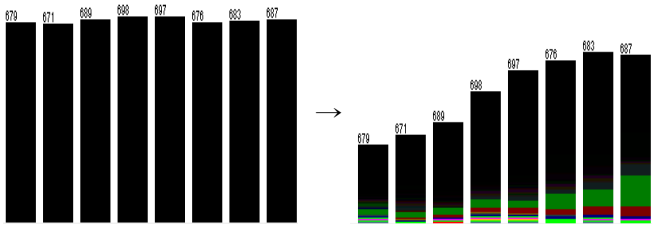

그림 1. Kernel Filtering 후 데이터의 변화
Fig. 1. Data Changes after Kernel Filtering

3.2 Kernel Filter 과정

본 논문에서 데이터를 Kernel Filter를 이용하여 전처리를 하였다. Kernel Filter는 새로운 데어터 집합을 출력하는데, 여기에는 이전과 동일한 개수의 인스턴스가 들어있다. 여기서 각 인스턴스의 값은 기존 인스턴스 한 쌍에 대해 커널 함수가 내린 평가 결과가 된다. 기본적으로 모든 값은 평균이 0에 가깝게 변환된다. 본 실험과정에서 Kernel Filter를 통한 전처리 과정을 통해 데이터들이 보기 좋은 집합 형태로 변환되었다.

3.3 실험과정

실험을 위한 도구로써 Waikato 대학교에서 개발된 WEKA v3.6.12을 사용하고[3], 사용된 데이터는 호주의 건강비용 지출이 기록된 health expenditure.arff 이다. 실험에 사용된 데이터는 총 5,480개를 사용한다.

연도에 따라서 사람들이 건강을 위해 건강비용 지출의 증가를 토대로 건강의 중요성을 인식하고 EM 알고리즘과 DBSCAN 알고리즘을 통해 비교 분석을 하고 Clustering을 통해 건강비용 지출을 예측한다.

3.4 실험 결과

실험은 health expenditure 데이터를 기반으로 real expenditure million 속성을 대상으로 선택한 탐색 알고리즘으로 앞에서 설명한 LinearRegression와 EM 알고리즘을 사용하였으며, 데이터 군집화에는 numcluster (군집 개수)의 값을 3으로 주어 수행하였다. 아래 그림에서는 연도에 따라서 DBSCAN의 실험 결과를 나타내었고 아래 그림에서 EM 알고리즘의 실험 결과를 나타내었다.

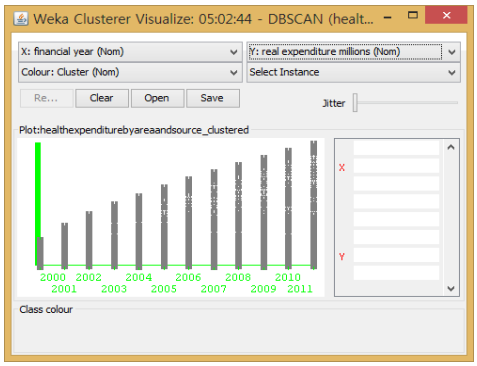

그림 2. DBSCAN 결과화면
Fig. 2. DBSCAN Results Screen

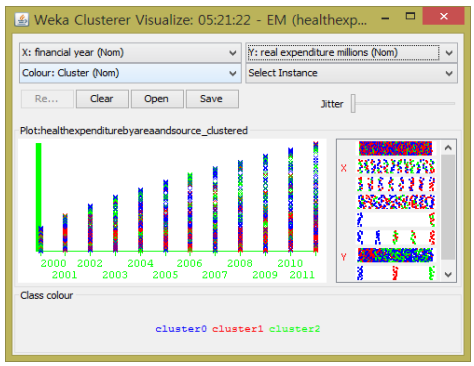

그림 3. EM 결과화면
Fig. 3. EM Results Screen

4. 평가 및 토론

4.1 실험 결과 평가

아래 그림은 군집을 3개로 가정해서 실험한 EM 알고리즘의 결과를 나타낸다. 아래 그림은 DBSCAN 알고리즘의 결과이다.

```
=== Model and evaluation on training set ===

Clustered Instances

0      2263 ( 41%)
1      2484 ( 45%)
2       733 ( 13%)


Log likelihood: -14.23446
```

그림 4. EM 알고리즘 결과
Fig. 4. EM Algorithm Results

아래 그림에서 보면 3개의 군집으로 나누어서 1번 군집에는 2263, 2번 군집에는 2484, 3번 군집에는 733개의 데이터가 그룹화 되어 들어가 있는 것을 볼 수 있다.

```
Clustered DataObjects: 5480
Number of attributes: 6
Epsilon: 0.9; minPoints: 6
Index: weka.clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase
Distance-type: weka.clusterers.forOPTICSAndDBScan.DataObjects.EuclideanDataObject
Number of generated clusters: 0
Elapsed time: 1.81
```

그림 5. DBSCAN 알고리즘의 결과
Fig. 5. DBSCAN Algorithm Results

아래 그림에서 보면 클러스터링이 진행된 데이터 오브젝트의 수가 나오고, attribute의 수, 사용자 정의 거리(Epsilon), 최소 포인트값 등이 나오게 된다. Health expenditure의 데이터로 실험한 결과를 아래 그림과 같이 나타낸다.

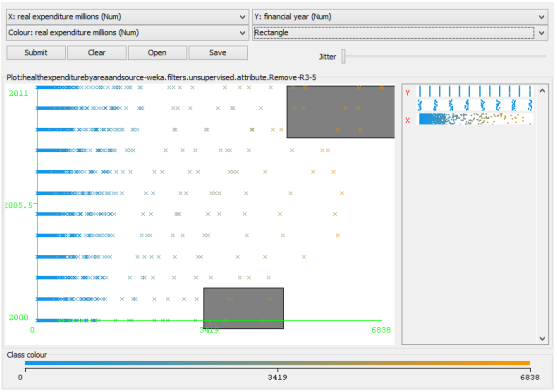

그림 6. Health Expenditure 결과
Fig. 6. Health Expenditure Results

위 그림을 살펴보면 2000년 초반과 2010년 초반의 최고 건강비용 지출을 비교할 수 있으며, DBSCAN 알고리즘과 EM 알고리즘의 성능을 동일한 데이터에 대하여 분석하였다. 이를 통하여 점점 건강에 대한 중요도가 커지고 그에 따른 건강비용 지출 또한 커지고 있음을 확인할 수 있다.

5. 결론

최근에는 건강의 인식이 더욱 중요시 되면서 다른 분야에서도 데이터들을 방대하게 가지고 있지만 의료분야, 안전분야 등 즉 건강에 관련된 많은 분야들의 데이터를 활발하게 사용되어지고 있다. 또한 건강에 관련된 데이터뿐만 아니라 다른 데이터들을 이용해 자료를 추출하여 사용 할 수 있는 좋은 정보를 얻기 위하여 데이터마이닝 분야와 빅데이터 분야가 더욱 각광받고 있다. 하지만 데이터가 방대함에 따라 데이터들을 완벽하게 다룰 수 있는 알고리즘이 개발되지 못한 현황이다. 따라서 본 논문 에서는 데이터마이닝 기법 중의 DBSCAN 알고리즘과 데이터마이닝 군집화 알고리즘 중의 하나인 EM 알고리즘의 성능을 동일한 데이터에 대하여 분석하였다. DBSACN과 EM 알고리즘에 따른 변화를 Health expenditure 실험데이터의 결과를 기반으로 분석하였다. 향후에

는 더욱 정확한 실험과 더욱 정확한 결과를 알아내기 위하여 더욱 정확한 데이터를 찾아내고 그에 맞는 알고리즘을 사용하고 그에 적용할 수 있는 다 수의 알고리즘을 찾아 복합적으로 적용하고 더욱 많은 데이터의 정보로 정확한 결과를 도출할 수 있게 연구할 것이다.

참 고 문 헌

〔국외 문헌〕

[1] Doddi, S., Achla Marathe, Ravi, S. S., and Torney, D. C. (2001), “Discovery of association rules in medical data”, Informatics for Health and Social Care, 26(1), 25-33.

[2] Hosking, J. R. M. and Wallis, J. R. (2005), Regional frequency analysis: an approach based on L-moments, Cambridge University Press.

[3] Kirchhoff, W. H. (2012), “LOGISTIC FUNCTION PROFILE FIT: A least-squares program for fitting interface profiles to an extended logistic functiona)”, Journal of Vacuum Science and Technology, A 30.5, 051101.

[4] Malefaki, S., Trevezas, S., and Limnios, N. (2010), “An EM and a stochastic version of the EM algorithm for nonparametric Hidden semi-Markov models”, Communications in Statistics-Simulation and Computation®, 39(2), 240-261.

[5] Palaniappan, S. and Awang, R. (2008), “Intelligent heart disease prediction system using data mining techniques”, 108-115.

[6] Witten, I. H. and Frank, E. (2005), Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.

정 용 규 (Yong Gyu Jung)

서울대학교, 연세대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고, 현재 을지대학교 의료IT마케팅학과 교수로 재직 중이다. ISO, UN의 전자거래분야 한국대표위원으로 활동하였으며, 의료정보, 전자무역, 물류유통 등에 Semantic Web, Process Modelling, ebXML 등의 표준 기술의 적용에 관심이 많다.

최 영 진 (Young Jin Choi)

성균관대학교 경영학과 경영학박사를 받았다. 1995년부터 2006년까지 한국정보화진흥원에서 수석연구원으로 정보시스템감리, 정보화성과평가, ITA/IT 거버넌스 등의 업무를 주로 수행하였다. 2006년부터는 을지대학교 의료경영학과 부교수로 재직 중에 있으며, 주요 관심분야는 IT 거버넌스, 의료정보, IT성과평가이다.

차 병 헌 (Byung Heun Cha)

한양대학교에서 의학박사 학위를 취득하였고 다양한 임상경험을 기반으로 2005년부터 을지대학교 교수로 재직하고 있다. 주요 관심분야로는 유전학과 관련한 유전자 분석과 노화 기전에 많은 연구와 관심을 기울이고 있다.

Clustering Analysis of Effective Health Spending Cost based on Kernel Filtering Techniques

Yong Gyu Jung* · Young Jin Choi** · Byeong Heon Cha***

ABSTRACT

As Data mining is a method of extracting the information based on the large data, the technique has been used in many application areas to deal with data in particular. However, the status of the algorithm that can deal with the healthcare data are not fully developed. In this paper, One of clustering algorithm, the EM and DBSCAN are used for performance comparison. It could be analyzed using by the same data. To do this, EM and DBSACN algorithm are changing performance according to the variables in Health expenditure database. Based on the results of the experimental data, We analyze more precise and accurate results using by Kernel Filtering. In this study, we tried comparison of the performance for the algorithm as well as attempt to improve the performance. Through this work, we were analyzed the comparison result of the application of the experimental data and of performance change according to expansion algorithm. Especially, Collects data from the various cluster using the medical record, it could be recommended the effective spending on medical services.

Keywords: Health Expenditure Data, EM, DBSCAN, Clustering, Regression Analysis

* Dept. of Medical IT and Marketing, Eulji University, ygjung@eulji.ac.kr
** Dept. of Healthcare Management, Eulji University, yuzin@eulji.ac.kr
*** Corresponding Author, Dept. of Biomedical Laboratory Science, Eulji University, jabogy@eulji.ac.kr