Journal of the Korean Data & Information Science Society
2017, **28**(2), 395–406

http://dx.doi.org/10.7465/jkdi.2017.28.2.395
한국데이터정보과학회지

시간단위 전력사용량 시계열 패턴의 군집 및 분류분석[†]

박다인[1] · 윤상후[2]

[1]대구대학교 일반대학원 통계학과 · [2]대구대학교 전산통계학과, 대구대학교 기초과학연구소

접수 2017년 2월 28일, 수정 2017년 3월 27일, 게재확정 2017년 3월 27일

요 약

전력 공급 시스템의 효율적인 운영을 위해 전력수요예측은 필수적이다. 본 연구에서는 군집분석과 분류분석을 이용하여 일 단위 시간별 전력수요량 시계열 패턴의 유형을 살펴보고자 한다. 전력거래소에서 수집된 2008년 1월 1일부터 2012년 12월 31일까지의 일 단위 시간별 전력수요량 데이터를 추세성분, 계절성분, 오차 성분으로 구성된 시계열 자료로 변환하여 사용하였다. 추세성분을 제거한 시계열 자료의 패턴을 구분하기 위한 군집 분석방법은 k-평균 군집분석 (k-means), 가우시안혼합모델 혼합 모델 군집분석 (Gaussian mixture model), 함수적 군집분석 (functional clustering)을 고려하였다. 주성분분석을 통해 24시간 자료를 2개의 요인로 축소한 후 k-평균 군집분석과 가우시안 혼합 모델, 함수적 군집분석을 수행하였다. 군집분석 결과를 토대로 2008년부터 2011년까지 총 4년간 데이터를 4가지 분류분석방법인 의사결정나무, RF (random forest), Naive bayes, SVM (support vector machine)을 통해 훈련시켜 2012년 군집을 예측하였다. 분석 결과 가우시안 혼합 분포기반 군집분석과 RF를 이용한 군집예측 결과의 성능이 가장 우수하였다.

주요용어: 군집분석, 기계학습, 분류분석, 전력수요량.

1. 서론

경제 성장은 전력 수요의 증가를 동반하므로 산업통상자원부와 한국에너지관리공단에서는 효율적인 에너지 관리를 위하여 에너지 저장 시스템 (Energy storage system; ESS)을 보급토록 하는 사업을 증가시키는 추세이다. 에너지 저장 시스템이란 과잉 생산된 전기를 저장하였다가 전기의 수요가 발생했을 때 전기를 공급하여 전력이용의 효율을 높여주는 시스템이다. 에너지 관리 시스템 (energy management system)은 에너지 저장 시스템의 핵심 부분으로 에너지 소비 패턴을 파악하여 에너지 수요량을 예측하여 적절한 수준의 에너지 발전이 되도록 시스템을 운영하고 관리한다. 에너지의 효율적인 관리는 에너지 소비 패턴을 어떻게 파악하느냐가 중요하다고 볼 수 있다.

전력수요 예측모형은 에너지 소비 패턴을 파악하는데 유용하다. 전력수요예측은 기간에 따라 크게 장기, 중기, 단기로 구분된다. 단기 예측은 하루에서 일주일 후 예측이 주를 이루며 중기 예측은 월단위에서 일년 후를 예측한다. 장기 예측은 전력수급계획을 위해 수년에서 수십년 후를 예측 기간으로 사용한다. 중기 전력수요예측은 전력의 수요관리와 전력생산을 위한 자원시장 운영의 기초자료로 연단위 주별 최대전력수요예측이 필요하다. 이러한 중기 전력수요예측에 가장 영향을 미치는 요인은 기상요인이

[†] 이 논문은 2016년도 대구대학교 학술연구비 지원에 의한 논문임.
[1] (38453) 경상북도 경산시 진량읍 대구대로201, 대구대학교 일반대학원 통계학과, 석사과정.
[2] 교신저자: (38453) 경상북도 경산시 진량읍 대구대로201, 대구대학교 전산통계학과, 조교수.
E-mail: statstar@daegu.ac.kr

다 (Wi와 Min, 2016). 기상요인은 불확실성을 내포하고 있지만 기상청에서 수개월에서 일년까지 지역별로 예측하고 있으므로 기상요인의 예측값을 반영한 전력수요예측이 가능하다.

전력수요의 패턴은 일반적으로 군집분석을 통해 파악된다. 전력수요 패턴의 연구로는 k-평균 군집분석 (Lim 등, 2013), 함수적 군집분석 (Yoon과 Choi, 2015), 계층적 군집분석 (Hwang 등, 2015)이 있다. 본 연구에서는 전력수요의 군집분석방법으로 k-평균 군집분석, 가우시안 혼합 모델 군집분석, 함수적 군집분석을 고려하였다.

군집분석을 통해 파악된 전력수요 패턴 결과에 대한 검증방법으로 본 연구에서는 분류분석을 이용하였다. 분류분석을 위한 종속변수는 군집분석 결과이고 독립변수는 기상요인, 휴일, 요일 등이다. 기상요인과 외부요인을 활용한 전력수요에 관한 연구는 데이터마이닝 (Kim 등, 2012), 일반화가법모형과 선형모형의 하이브리드 모형 (Cho 등, 2013), 시간단위 기온을 이용한 시계열모형 (Kang 등, 2016), 퍼지모형 (Song 등, 2005) 등이 제안되었다.

본 논문의 구조는 다음과 같다. 2절에서는 본 연구에서 사용하고자 하는 군집분석과 분류분석에 관한 방법론을 살펴보고 3절에서는 연구자료와 분석결과를 확인한다. 마지막으로 4절에서는 연구 결과에 대한 요약정리 및 한계점을 정리한다.

2. 군집분석 및 분류분석

이번 절에서는 전력수요 패턴을 파악하기 위해 본 연구에서 고려한 군집분석 및 분류분석 방법에 대해 설명하고자 한다. 일단위 전력수요 패턴을 파악하기 위해 k 평균 군집분석, 가우시안 혼합 모델 군집분석, 함수적 군집분석이 고려되었다. 군집분석 결과는 기상요소, 휴일, 요일을 기반으로 한 분류분석에 이용된다.

2.1. k-평균 군집분석 (k-means clustering)

MacQueen (1967)에 따르면 k-평균 군집분석은 군집의 수 (k)를 미리 정하여 각 개체가 어느 군집에 할당되는지 분석하는 상호 배반적 군집방법이다. k-평균 군집분석의 절차는 다음과 같다.

(1) 전체 개체를 k개의 군집으로 초기화한다.

(2) 각 군집별 중심점을 찾는다.

(3) 모든 개체와 각 군집의 중심점과의 유클리드 거리를 계산하여 개체와의 거리가 가장 가까운 중심점에 대응하는 군집으로 배정 한다.

(4) 모든 개체가 다른 군집으로 재배정되지 않을 때까지 (2)∼(4) 과정을 반복한다.

여기서 p차원 상의 두 관찰 값 $\mathbf{X}$=(x_1,x_2,$\cdots$,x_p)와 $\mathbf{Y}$=(y_1,y_2,$\cdots$,y_p)간의 유클리드 거리는 $d(\mathbf{X},\mathbf{Y}) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_p-y_p)^2} = \sqrt{(\mathbf{X}-\mathbf{Y})^t(\mathbf{X}-\mathbf{Y})}$로 정의된다. 유클리드 거리를 이용하고 군집의 중심을 그 군집에 속한 관측치들의 평균으로 정의하므로 모든 변수가 연속형 변수여야 한다. k-평균 군집분석은 적절한 k의 결정에 어려움이 있고 이상치 데이터가 있을 경우 초기값에 민감하다는 단점이 있다.

2.2. 가우시안 혼합 모델 군집분석 (Gaussian mixture model clustering)

Scott와 Symons (1971)가 제안한 가우시안 혼합 모델은 여러 개의 가우시안 함수를 사용하여 복잡한 형태의 자료를 군집화 한다. 모집단이 G개의 군집으로 구성된다면 k번째 군집에 속한 p차원 관측벡터 $\boldsymbol{x}$의 밀도함수는 $f_k(\boldsymbol{x},\boldsymbol{\theta})$라고 가정한다. 이때 우도함수는 다음과 같다.

$$L(\boldsymbol{\theta},k) = \Pi_{i=1}^{n} f_k(x_i;\theta_k) \tag{2.1}$$

우도함수의 최대화를 통해 $\boldsymbol{\theta}$가 결정되고 각 개체는 우도 계산을 통한 사후확률값의 크기가 가장 큰 k에 배정된다. 여기서 혼합모형 (mixture model)을 고려하면 다음과 같다.

$$L_{min}(\boldsymbol{\theta}, k) = \Pi_{i=1}^{n} \Sigma_{k=1}^{G} \tau_k f_k(x_i; \theta_k) \tag{2.2}$$

여기서 τ_k는 관측벡터가 k번째 군집에 속할 사전확률이며 $\tau_k \geq 0$이고 $\Sigma_{k=1}^{G} \tau_k = 1$을 만족한다. 이제 관측벡터가 다변량 정규분포를 따른다고 가정하는 가우시안 혼합모형을 고려해보자. $f_k(\boldsymbol{x}, \boldsymbol{\theta})$는 평균벡터 $\boldsymbol{\mu_k}$와 공분산행렬 Σ_k를 갖는 k번째 군집의 다변량정규분포의 밀도함수는 다음과 같다.

$$\Phi_k(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_k) = (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(x_i - \boldsymbol{\mu_k})^t \Sigma_k^{-1} (x_i - \boldsymbol{\mu_k})\right) \tag{2.3}$$

가우시안 혼합모델의 군집들은 타원형을 이루고, 평균 $\boldsymbol{\mu_k}$를 중심으로 분포되어 있다. 각각의 공분산 행렬은 $\Sigma_k = \lambda_k D_k A_k D_k^t$의 형태로 모수화 된다. Σ_k의 주요성분은 D_k에 의해 결정되고, A_k에 따라 밀도의 윤곽이 나타난다. 여기서 D_k는 고유벡터의 직교행렬이며 군집의 방향을 결정한다. A_k는 Σ_k의 고유값에 비례적으로 취하는 대각행렬로 군집의 형태를 결정한다. 마지막으로 λ_k는 스칼라로 군집의 크기를 결정한다.

군집 개수 G가 정해지면 가능한 군집 ($1 \leq k \leq G$)에 대해 EM알고리즘에 의해 추정된다. E 단계에서는 주어진 조건에서 관측벡터가 각 군집에 속할 확률을 구하고 M 단계에서는 주어진 상황에서의 모수가 추정된다. 가우시안 혼합모델의 군집분석을 R의 'mclust' 패키지를 통해 수행되었다 (Fraley 등, 2016).

2.3. 함수적 군집분석 (functional clustering)

함수적 군집분석은 함수적 자료 (functional data)를 군집화하는 방법으로 본 연구에서는 Ma 등 (2006)이 제안한 시간 함수적 자료를 위한 군집모형을 고려되었다. 함수적 자료 분석에서는 관측벡터 $\boldsymbol{x}$가 평활함수로 표현되며 i번째 개체 (곡선) 값은 다음과 같이 표현된다.

$$Y_i = f(x_i) + \epsilon_i, \ i = 1, \cdots, n \tag{2.4}$$

여기서 n은 개체의 수를 나타내고 ϵ_i는 평균이 0이고 분산이 σ^2인 가우시안분포를 따른다. 일반적인 평활함수는 잔차제곱합 (residual sum of squares)을 최소화하는 함수이다.

$$RSS = \Sigma_{i=i}^{n} [y_i - f(x_i)]^2 \tag{2.5}$$

함수적 군집분석는 혼합가우시안 분포를 가정하여 penalized 로그-우도함수를 최대로 하는 모수를 추정하기 위해 EM알고리즘과 Generalized Cross-Validation (GCV)를 이용한다. 함수적 군집분석은 R의 'MFDA'패키지를 이용하였으며, 구체적인 군집분석방법은 Yoon과 Choi (2015)을 참고바란다.

2.4. 의사결정나무 (decision tree)

의사결정나무는 데이터 분리 즉, 노드 분리로써 관심 대상을 소집단으로 분류하거나 예측하는 분석방법이다. 의사결정나무 구조는 뿌리노드로부터 시작하고 핵심은 노드 분리에 있다. 노드분리는 분리대상인 노드 M을 자녀 노드 C_1과 C_2로 나누는 것이다. x중 하나의 x_j와 어떤 값 k_j를 선택하여, $x_j \leq k_j$인 개체는 노드 C_1에 넣고 $x_j > k_j$인 개체는 노드 C_2에 넣는다. 변수 x_j와 분리 값 x_j의 선택은 노드의 불순도 (impurity)로 결정한다. 의사결정나무 모형은 의사결정나무 형성-가지치기-타당성 평가-해석 및 예측으로 수행된다. 의사결정나무 형성 단계에서는 데이터 분석의 목적과 구조에 따

라 적절한 분리기준과 정지 기준을 지정하여 의사결정나무를 형성한다. 가지치기 단계에서는 분류오류를 크게 하는 위험이 높거나 적절하지 않은 추론규칙을 가지고 있는 가지를 제거한다. 타당성평가 단계에서는 이익도표, 위험도표, 교차타당성 등을 이용하여 의사결정나무를 평가한다. 본 연구에서는 R의 'rpart'패키지를 이용하여 분류와 회귀에 적용되는 CART (classification and regression tree)를 수행하였다 (Therneau 등, 2015).

2.5. 랜덤포레스트(random forests)

랜덤포레스트는 여러 개의 의사결정나무들로 구성된 모형이다 (Breiman, 2001). 랜덤포레스트는 붓스트랩 표본을 다수 생성하고 이를 의사결정나무 모형을 적용하여 그 결과를 종합하는 앙상블 방법 (ensemble methods)이다. 랜덤포레스트에서 생성된 의사결정나무모형 간 상관이 낮을수록 예측오차가 작으므로 무작위로 생성된 붓스트랩 표본의 의사결정나무모형들이 서로 독립일수록 예측오차가 작아진다. 또한, 의사결정나무 수가 많아도 랜덤포레스트는 과적합 하지 않는다는 장점이 있다 (Park, 2016). 본 연구에서는 R의 'randomForest' 패키지를 이용하였다 (Liaw와 Wiener, 2002).

2.6. 나이브 베이즈 (naive Bayes)

나이브 베이즈는 조건부 확률모델로 분류될 개체들은 n개의 설명변수를 나타내는 벡터 x로 표현되며, 나이브 베이즈 분류기는 이 벡터를 이용하여 k개의 가능한 확률적 결과들을 다음 식과 같이 할당한다.

$$p(C_k|x_1,,x_n) = \frac{p(C_k)p(\boldsymbol{x}|C_k)}{p(\boldsymbol{x})} \tag{2.6}$$

만약 독립성 가정하에서 그룹 의 조건부 분포는 아래와 같다.

$$p(C_k|x_1,\cdots,x_n) = \frac{1}{Z}p(C_k)\prod_{i=1}^{n}p(x_i|C_k) \tag{2.7}$$

여기서 $Z = p(x)$로 $x_1,\cdots,x_n$에만 의존하는 규모요소이다. 새로운 입력벡터는 가장 가능성 높은 집단에 속하는 것으로 에 C_k에 대해서 다음 식을 통해 최대 확률을 갖는 그룹 k를 찾아낸다.

$$\hat{y} = argmax_{k\in(1,\cdots,k)}p(C_k)\prod_{i=1}^{n}p(x_i|C_k) \tag{2.8}$$

2.7. 서포트 벡터머신 (support vector machine, SVM)

서포트 벡터머신은 서포트 벡터를 통해 학습데이터에 대한 오차를 최소화하는 기계학습방법이다. 그룹을 구성하는 설명변수들을 선형으로 분리한다고 하면, SVM은 하나의 집단과 다른 집단을 분류하는 최적의 분리경계면 (optimal boundary hyperplane)을 찾는 것이다.

선형 분리가 가능한 경우 최적의 분리경계면은 서포트 벡터들의 중간점을 통과하는 것으로 정의된다. 찾고자하는 선형 분류함수를 $f(x) = w^tx + b$ 라 하자. $f(x) > 0$이냐 $f(x) < 0$이냐에 따라 서로 다른 두 개의 그룹으로 분류한다. 제약조건완화에 패널티를 부과하고 라그랑지 승수를 이용하면 해를 구할 수 있다. $y_i = 1$인 x_i에 대하여 $w^tx_i - b \geq 1 - \xi_i$, $y_i = -1$ 인 x_i에 대하여 $w^tx_i - b \leq -1 + \xi_i$를 만족하는 $\frac{1}{2}||w||^2 + C\Sigma_{i=1}^{n}\xi_i$을 최소화 하자. 여기서 $\xi_1 \geq 0,\cdots,\xi_n \geq 0$은 조건완화를 위한 여분 (slack)이고 $C > 0$은 여분에 부과하는 단위비용 (unit cost)이다.

선형분리가 불가능할 경우 커널방법론을 이용한다. 자료를 특성 공간으로 매핑하여 매핑된 특성값 $\phi(x_i)$에 대해 선형 서포트 벡터 분류기를 적용하면 다음과 같은 최적화 문제를 얻는다.

$$\min_{\alpha} \left(\frac{1}{2} \sum_{i=1}^{2} \sum_{i=j}^{n} y_i y_j \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle - \sum_{i=1}^{n} \right) \tag{2.9}$$

위의 식에서 구체적인 ϕ를 모르더라도 그 내적 값만 계산 할 수 있으면 분류 함수를 얻을 수 있다. 즉 커널함수 $K(x, x^t) = \phi(x), \phi(x^t)$만 알면 충분하다. 최적 경계는 양 그룹에 대한 여백 경계(margin)의 중간 위치에서 결정되며, 서포트 벡터는 예백경계의 반대쪽에 놓이거나 여백 경계 바로 위에 놓인 관측치들을 지칭한다.

본 연구에서는 R의 'e1071'패키지를 통해 나이브 베이즈와 서포트 벡터머신를 이용한 분류를 수행하였다 (Meyer 등, 2015; Dimitriadou 등, 2005)

3. 자료분석

본 연구 자료는 전력거래소에서 수집된 2008년 1월 1일부터 2012년 12월 31일까지 시간단위 전력수요량 자료가 사용되었다. 시간에 따른 전력수요량의 시계열 그림은 Figure 3.1이다. 일반적인 시계열자료와 유사하게 추세성과 계절성이 존재하고 있다. 시간의 흐름에 따라 자료의 변동 폭이 증가하므로 로그변환을 실시하고 추세성이 전력수요량의 유형을 파악하는데 영향을 미칠 수 있으므로 추세성을 제거하여 연구를 수행하였다. 로그변환과 추세성이 제거된 자료는 시간과 무관한 계절성만이 반영된 자료이다 (Figure 3.1).

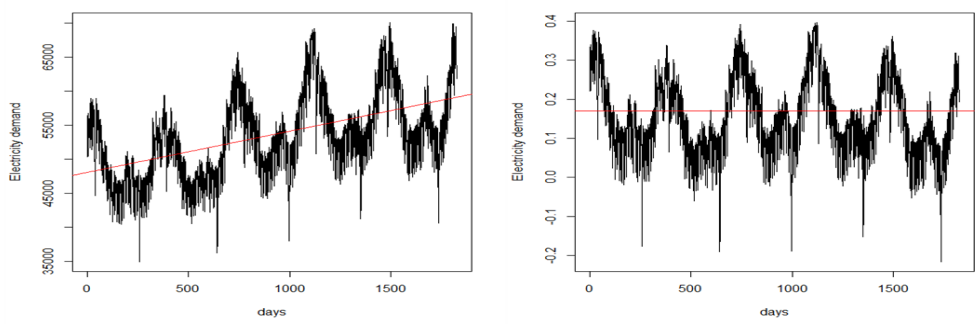

Figure 3.1 The time series plot of electricity demands (left : raw data, right: after eliminating time trend)

Song 등 (2005)과 Lim 등 (2013)에 의하면 전력 수요 패턴은 요일, 휴일, 날씨, 사회적 이벤트 등에 의해 다양한 특징을 보인다. 24시간 전력수요 패턴은 주중, 주말, 특수일, 기타로 크게 4가지로 나눌 수 있다. 주중은 다시 월요일과 화요일-금요일로 나뉜다. 월요일은 주말에 쉬는 기업이나 공장의 영향으로 오전시간대의 전력수요가 현저히 낮아 다른 평일과는 전력 수요 패턴이 달라지기 때문이다. 주말에는 토요일과 일요일이 해당하는데 평일보다 낮은 전력 수요량을 보인다. 특수일은 삼일절이나 어린이날과 같은 나라에서 지정한 공휴일과 선거일과 같은 사회적 이벤트가 있는 날이다. 기타에는 여름휴가기간, 특수일 전후일, 징검다리요일, 설날이나 추석과 같이 공휴일이 긴 기간들이 포함된다.

하루는 24시간으로 구성되어 있으므로 각 시간을 하나의 개별 독립변수로 간주한다면 하루의 전력수요패턴은 24개의 독립변수를 이용하여 군집화 할 수 있다. 군집분석은 모든 독립변수에 동일한 가중치를 부여하므로 본 연구에서는 요인분석을 통해 서로 독립인 요인들을 추출하여 접근하였다. 요인분석 결과 하루 24시간의 전력수요량은 2개의 요인으로 정리 된다. 첫 번째 요인은 일과시간인 오전 9시부터 오후 10시까지 적재값이 높으므로 낮요인, 두 번째 주성분은 일과이후인 밤시간대에 적재값이 높아 밤요인으로 지칭하였다. 낮요인은 전체분산의 약 57.9%를 설명하고 밤요인은 전체분산의 약 39.3%를 설명하여 두 개의 요인으로 전체분산의 약 97.2%가 설명된다.

Table 3.1 The result of factor analysis

Hours of day	Factor1	Factor2
0h ∼ 1h	0.221	0.962
1h∼2h	0.220	0.961
2h∼3h	0.227	0.961
3h∼4h	0.227	0.968
4h∼5h	0.220	0.975
5h∼6h	0.230	0.967
6h∼7h	0.364	0.914
7h∼8h	0.619	0.760
8h∼9h	0.826	0.526
9h∼10h	0.905	0.389
10h∼11h	0.930	0.340
11h∼12h	0.945	0.300
12h∼13h	0.944	0.274
13h∼14h	0.971	0.203
14h∼15h	0.980	0.166
15h∼16h	0.983	0.155
16h∼17h	0.983	0.168
17h∼18h	0.957	0.253
18h∼19h	0.916	0.351
19h∼20h	0.928	0.328
20h∼21h	0.939	0.292
21h∼22h	0.896	0.395
22h∼23h	0.694	0.672
23h∼24h	0.557	0.784
SS loadings	13.892	9.440
Proportion Variance	0.579	0.393
Cumulative Variance	0.579	0.972

두 개의 요인을 통해 k평균 군집분석과 가우시안 혼합 모델 군집분석을 실시한 결과는 Figure 3.2이다. 함수적 군집분석의 경우 차원축소 없이 24개의 시간별 자료를 평활 스플라인 (smoothing spline)을 통한 부드러운 함수자료로 간주하여 분석하였다. 군집이 Figure3.2의 오른쪽으로 위치해 있을수록 낮시간 전력수요량이 높고 위쪽으로 위치해 있을수록 밤 시간에 전력수요량이 높다. k-평균 군집은 각 군집의 중심점을 중심으로 데이터들이 원형으로 군집되어 있으나 가우시안 혼합모델 군집의 경우 데이터들이 타원형으로 군집되어 있다. 함수적 군집방법은 k-평균과 가우시안혼합모델군집방법이 혼합된 형태로 보인다.

군집분석에서 적절한 군집수를 결정하는 것은 군집화 결과의 타당성에 전제가 되는 매우 중요한 문제이다. 본 연구에서는 군집의 수를 7부터 10까지 고정하여 계산하였다. 가우시안 혼합 모델 군집분석과 함수적 군집분석은 BIC를 기준으로 최적 군집 수를 결정되고 k-평균 군집분석은 연구자의 경험이나 정보기준 접근법 등으로 결정된다. 최적 군집 수를 결정하기 위해 본 연구에서는 외부요인을 통한 분류분석 결과를 활용하였다.

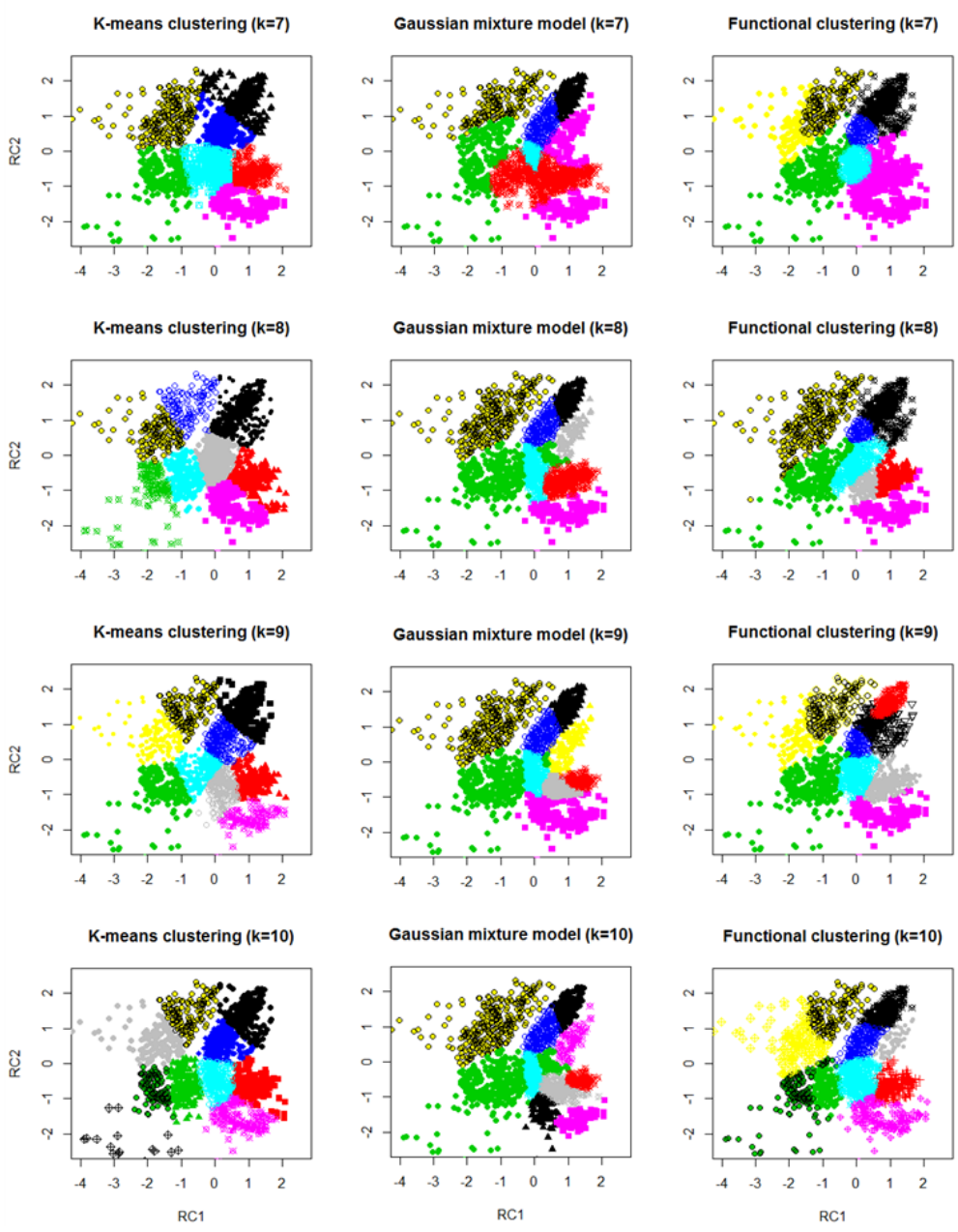

Figure 3.2 The result of cluster analysis

최적 군집 수 결정을 위한 분류분석은 다음과 같이 수행하였다. 먼저 2008년 1월 1일부터 2012년 12월 31일까지 군집분석을 실시하여 각 요일별 군집을 구분하였다. 전력수요예측이 목적이므로 2008년 1월 1일부터 2011년 12월 31일까지 수집된 기상요인, 휴일, 요일 데이터와 군집분석 결과를 훈련시켰

다. 훈련된 모형을 통해 2012년 1월 1일부터 2012년 12월 31일까지 수집된 기상요인, 휴일, 요일 데이터로 군집 번호를 예측하여 실제 관측된 군집분석 결과와 비교를 통해 예측성능을 평가하였다.

2008년부터 2011년까지의 훈련자료 (training data)를 이용한 분류분석 예측정확도 결과는 Figure 3.3이다. 예측성능은 랜덤포레스트, 의사결정나무, 서포트벡터머신, 나이브베이즈 순으로 랜덤포레스트의 결과가 가장 우수하지만 과대적합 (overfitting)일 가능성이 있기때문에 검증자료의 결과를 살펴보았다.

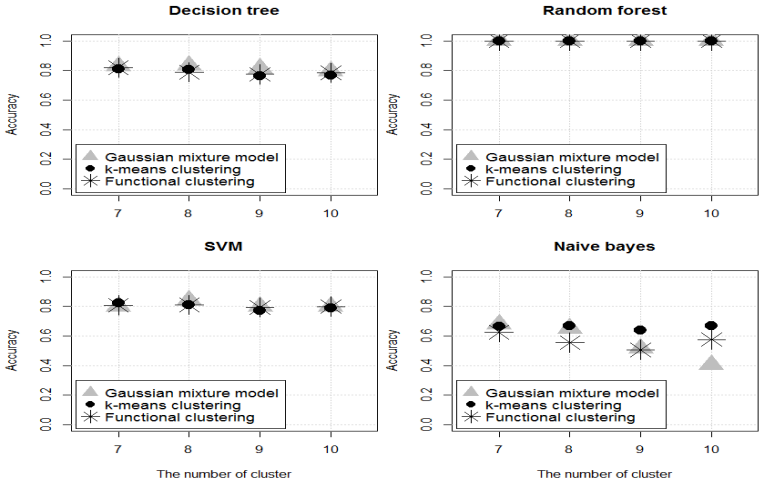

Figure 3.3 The prediction accuracy of classification (modeling data)

검증자료(test data)의 분류분석 예측정확도는 Table 3.2와 Figure 3.4이다. 분석결과 군집수가 8개이고 가우시안혼합모델을 이용한 군집분석을 수행하고 랜덤포레스트로 분류분석을 수행했을 경우의 예측 정확도가 86.6%로 가장 우수하였다. 예측정확도의 추세를 보면 군집의 수가 증가한다고 예측정확도가 증가하지 않고 오히려 감소하는 경향을 확인할 수 있다. 본 연구결과에서는 생략하였지만 가우시안 혼합 모델 군집분석의 최적 군집 수는 13이였고 함수적 군집분석에서는 군집의 수가 증가할수록 BIC가 낮아져서 적절한 군집의 수를 발견하지 못하였다. 하지만 본 연구의 분류분석 결과를 토대로 최적 군집 수를 결정한다면 최적 군집 수는 8이다. 예측 정확도가 가장 높은 랜덤포레스트의 중요 변수를 Figure 3.5에서 보여준다. 전력수요 패턴의 군집에 영향을 미치는 변수의 중요 순서는 요일, 평균기온, 최저기온, 휴일 여부, 최고기온의 순으로 나타났다.

Table 3.2 The result of accuracy (test data)

Classification	Clustering	The number of clusters			
		7	8	9	10
Decision tree	K-means	0.795	0.656	0.754	0.680
	Gaussian mixture	0.762	0.822	0.773	0.768
	Functional	0.795	0.746	0.730	0.732
Random forest	K-means	0.817	0.730	0.790	0.746
	Gaussian mixture	0.754	0.866	0.792	0.765
	Functional	0.779	0.760	0.768	0.738
Naive bayes	K-means	0.626	0.585	0.571	0.511
	Gaussian mixture	0.593	0.678	0.601	0.628
	Functional	0.628	0.568	0.533	0.527
SVM	K-means	0.751	0.645	0.710	0.615
	Gaussian mixture	0.683	0.724	0.661	0.689
	Functional	0.700	0.637	0.672	0.656

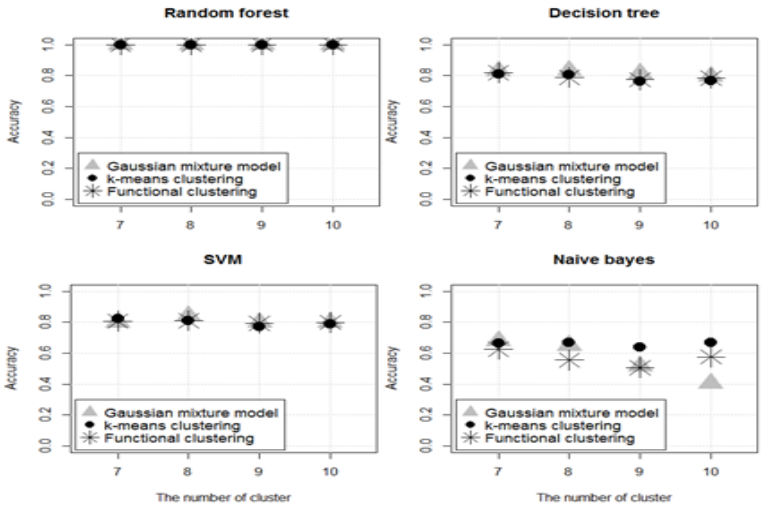

Figure 3.4 The prediction accuracy of classification (test data)

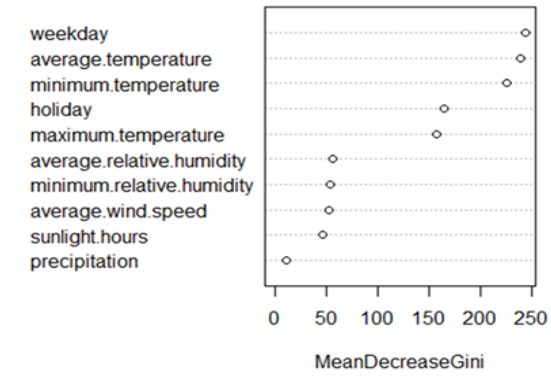

Figure 3.5 The important variable plot of random forest

군집분석과 분류분석을 통해 얻어진 전력수요패턴을 구분하는 최적 방법은 군집 수 8가 개인 가우시안 혼합 모델 군집분석이다. 군집 별 전력수요 패턴의 특징을 파악하기 위해 시간 단위 평균 전력수요량 커브와 95%구간을 그리면 Figure 3.6이다. 1군집은 난방의 영향으로 낮과 밤의 전력수요가 높고 설날과 크리스마스를 포함한 겨울철 휴일, 공휴일이다. 2군집은 난방으로 인한 아침시간대의 전력수요량이 높은 겨울철 월요일이다. 3군집은 난방의 영향으로 낮과 밤의 전력수요가 높은 한겨울 평일이다. 4군집은 봄, 가을, 초겨울의 평일이며 5군집은 봄과 가을의 평일 또는 여름철 주말을 대표하고 있다. 4군집과 5군집은 밤 기온에 따른 난방과 냉방에 유무에 따라 다르게 군집되었다. 6군집은 봄, 여름, 가을의 주말 및 공휴일로 어린이날이나 선거일과 같은 특수일을 포함하고 있다. 7군집은 오후부터 전력수요량이 급증하는 여름철 월요일이고 8군집은 일과시간의 전력수요량이 높은 여름철 평일을 대표한다. 각 군집의 패턴을 토대로 군집 유형을 요일, 휴일, 계절별로 정리하면 Table 3.3이다.

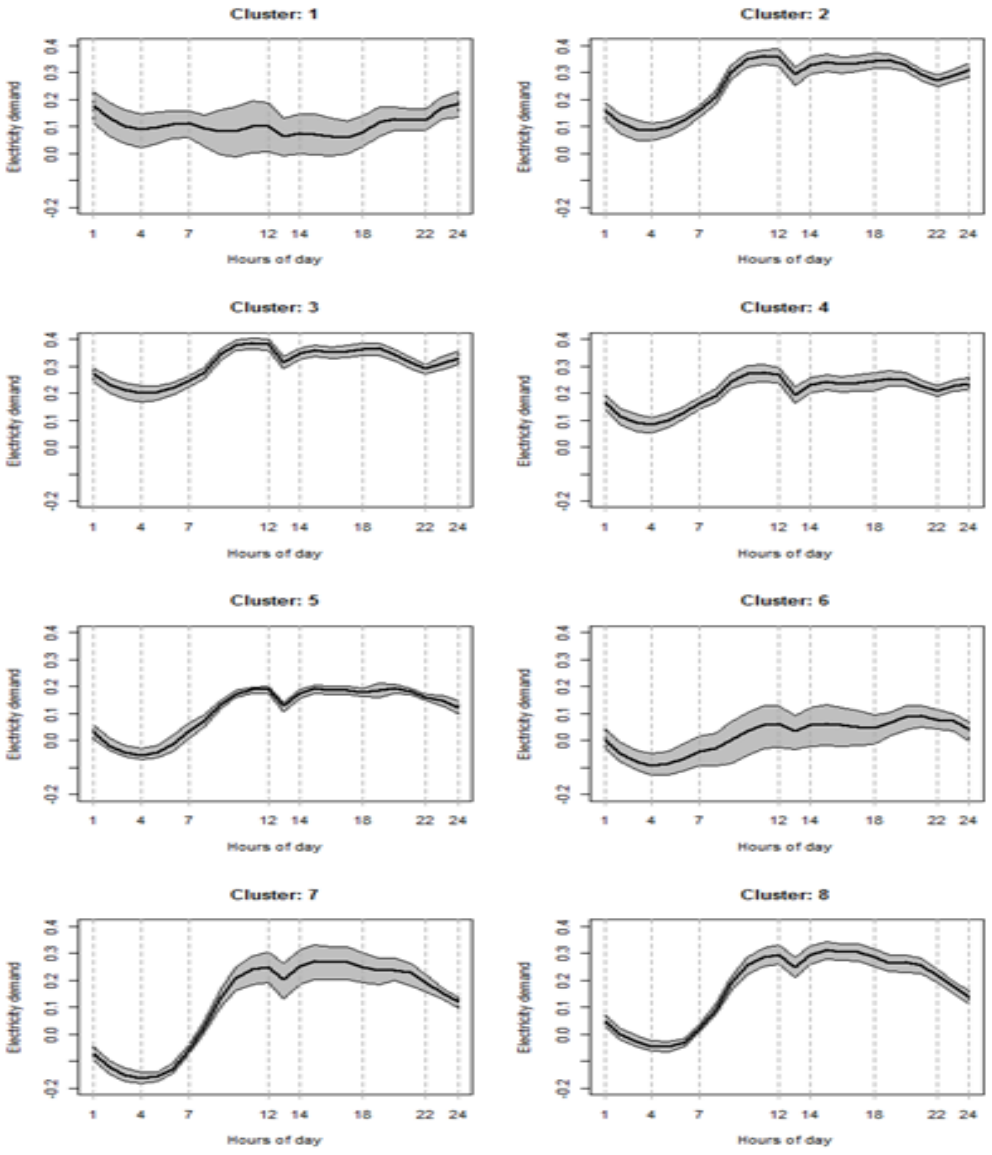

Figure 3.6 The mean curve and 95% interval for each cluster)

Table 3.3 The Characteristic of cluster based on month and weekday

Month	1	2	3	4	5	6	7	8	9	10	11	12
Monday		2			8			7			8	2
Tuesday to Friday		3		4		5		8		5	4	3
Holiday or Saturday or Sunday			1					6				1

4. 결론 및 고찰

Hwang 등 (2015)은 계층적 군집분석과 외부요인과의 관계를 통해 8개의 건물 부하패턴을 파악하였으나 군집 수의 결정방법은 명확하게 제시하지 못하고 있다. 본 연구는 전력수요 유형을 나누기 위해 군집분석을 사용하였고, 최적의 군집수의 결정을 위해 분류분석을 이용하였다. 분류분석을 위해 기온, 강수량, 풍속, 습도, 일조량, 휴일여부와 같은 외부요인이 고려되었다. 의사결정나무, 랜덤포레스트, 서포트벡터머신, 나이 브베이즈 4가지 분류방법으로 전력수요 패턴을 분류한 결과 랜덤포레스트 방법이 가장 우수하였다. 기상청에서는 중장기 기상예측정보를 지역별로 제공하므로 휴일과 요일이 주어진다면 하루 동안의 평균 전력수요패턴을 랜덤포레스트 방법을 통해 예측할 수 있다. 분류분석의 정확도를 기준으로 선정된 최적 군집 수는 8개이고 가우시안 혼합 모델을 이용한 군집분석이 전력수요 패턴의 군집에 가장 적절하다.

References

Breiman, L. (2001). Random forests. *Machine learning*, **45**, 5-32.

Cho, H., Goude, Y., Brossat, X. and Yao, Q.(2013). Modeling and forecasting daily electricity load curves: A hybrid approach. *Journal of the American Statistical Association*, **108**, 7-21.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A. (2005). *Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.5-7, `http://CRAN.R-project.org/.`

Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B. and Fop, M. (2016). mclust: Normal mixture modelling for model-based clustering, classification, and density estimation, `http://CRAN.R-project.org/package=mclust.Rpackageversion,5.`

Hwang, H. M., Lee, S. H., Park, J. B., Park, Y. G., and Son, S. Y. (2015). Load forecasting using hierarchical clustering method for building. *Journal of the Korean Institute of Illuminating and Electrical Installation Engineers*, 59-65.

Kang, D. H., Park, J. D. and Song, K. B. (2016). 24-Hour load forecasting for anomalous weather days using hourly temperature. *The Transactions of The Korean Institute of Electrical Engineers*, **65**, 1144-1150.

Kim, C. H., Koo, B. G. and Park, J. H. (2012). Short-term electric load forecasting using data mining technique. *Journal of Electrical Engineering & Technology*, **7**, 807-813.

Liaw, A, and Wiener, M. (2002). Classification and regression by randomForest. *IR news*, **2**, 18-22

Lim, J. H., Kim, S. Y., Park, J. D. and Song, K. B. (2013). Representative temperature assessment for improvement of short-term load forecasting accuracy. *Journal of the Korean Institute of Illuminating and Electrical Installation Engineers*, **27**, 39-43.

Ma, P., Castillo-Davis, C. I., Zhong, W. and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, **34**, 1261-1269.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C. and Lin, C. C. (2015). Package 'e1071'. *The Comprehensive R Archive Network*, Available at `https://cran.r-project.org/web/packages/e1071/e1071.pdf.`

Park, C. (2016). A simple diagnostic statistic for determining the size of random forest. *Journal of the Korean Data & information Science Society*, **27**, 855-863.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387-397.

Song, K. B., Baek, Y. S., Hong, D. H., and Jang, G. (2005). Short-term load forecasting for the holidays using fuzzy linear regression method. *IEEE transactions on power systems*, **20**, 96-101.

Therneau, T., Atkinson, B., Ripley, B., and Ripley, M. B. (2015). Package 'rpart', Available online `cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf.`

Wi, Y. M. and Min, Y. K. (2016). Weekly peak load forecasting using weather stochastic model and weather sensitivity. *The Transactions of the Korean Institute of Electrical Engineers*, **64**, 41-47.

Yoon, S. H. and Choi, Y. J. (2015). Functional clustering for electricity demand data: A case study. *Journal of the Korean Data & information Science Society*, **26**, 885-894.

Journal of the Korean Data & Information Science Society
2017, **28**(2), 395–406

http://dx.doi.org/10.7465/jkdi.2017.28.2.395
한국데이터정보과학회지

Clustering and classification to characterize daily electricity demand[†]

Dain Park[1] · Sanghoo Yoon[2]

[1]Department of Statistics, Daegu University
[2]Department of Statistics and Computer Science, Daegu University & Institute of Basic Science, Daegu University

Received 28 February 2017, revised 27 March 2017, accepted 27 March 2017

Abstract

The purpose of this study is to identify the pattern of daily electricity demand through clustering and classification. The hourly data was collected by KPS (Korea Power Exchange) between 2008 and 2012. The time trend was eliminated for conducting the pattern of daily electricity demand because electricity demand data is times series data. We have considered k-means clustering, Gaussian mixture model clustering, and functional clustering in order to find the optimal clustering method. The classification analysis was conducted to understand the relationship between external factors, day of the week, holiday, and weather. Data was divided into training data and test data. Training data consisted of external factors and clustered number between 2008 and 2011. Test data was daily data of external factors in 2012. Decision tree, random forest, Support vector machine, and Naive Bayes were used. As a result, Gaussian model based clustering and random forest showed the best prediction performance when the number of cluster was 8.

Keywords: Classification analysis, Cluster analysis, Electricity demand, Machine learning.

[†] This research was supported by the Daegu University Research Grant 2016.
[1] Master's course, Department of Statistics, Daegu University, Gyeongsan 38453, Korea.
[2] Corresponding author: Department of Statistics and Computer Science, Daegu University & Institute of Basic Science, Daegu University, Gyeongsan 38453, Korea. E-mail: statstar@daegu.ac.kr