기대치-최대화 군집 알고리즘과 출현 패턴 마이닝을 이용한 전력 소비 패턴 분석[1)]

박진형*, 이헌규*, 신진호**, 류근호*, 김희석***
*충북대학교 데이터베이스/바이오인포메틱스 연구실
**한국전력연구원 전력 정보 기술 그룹
***청주대학교 전자정보공학부
e-mail:{neozean, abolkog, hglee, khryu}@dblab.chungbuk.ac.kr
**jinho@kepri.re.kr
***khs8391@cju.ac.kr

Power Consumption Patterns Analysis Using Expectation-Maximization Clustering Algorithm and Emerging Pattern Mining

Jin Hyoung Park*, Heon Gyu Lee*, Jin-Ho Shin**, Keun Ho Ryu*, Hiseok Kim***
*Database/Bioinformatics Laboratory, Chungbuk National University
**Power Information Technology Group, Korea Electric Power Research Institute
***Division of Electronics & Information Engineering, College of Science & Engineering, Cheongju University

요 약

전력 회사의 효율적인 운용과 전력 시장에서의 경쟁을 위하여 고객의 전력 소비 패턴 분석 및 정확한 예측이 이루어져야 한다. 이를 위해서 이 논문에서는 원격 검침 시스템에 의한 전국의 고압 고객 데이터를 대상으로 고객의 전력 소비 패턴을 정확히 예측할 수 있는 마이닝 기법을 제안하였다. 먼저, 국내 계약종별 고객 특성에 맞는 부하 패턴의 정확한 구별을 위한 9가지의 특징 벡터를 추출하였고, 기대치-최대화 군집화 알고리즘을 사용하여 고객의 34개 대표 부하프로파일을 생성하였다. 마지막으로 추출된 특징 벡터로부터 각 대표 프로파일에 대한 출현 패턴 기반의 분류 모델을 구성하여 고객의 전력 소비 패턴을 분류하였다. 국내 원격 검침 시스템에 의해 측정된 총 3,895명의 고압 고객 데이터에 대한 실험 결과 약 91%의 분류 정확성을 보였다.

1. 서론

전력 산업에 있어서 전력 공급자의 효율적인 운영과 계획을 위하여 고객의 전력 사용 패턴의 특성 분석 및 예측 기술은 중요한 요소로 작용하고 있다. 최근에 원자재 가격 상승과 국가차원에서의 전기세 인상 규제 정책들로 인하여, 고객 수요에 맞는 전력 생산이 중요한 요소로 부각되면서 고객의 전력 사용 추이를 분석하고 예측하기 위해서 많은 연구들이 이루어지고 있다. 고객 전력 부하 패턴 분석 및 예측을 위하여 현재 통계적인 방법[1][2]과 데이터 마이닝 기법[3][4]이 적용되고 있다. 부하 예측의 정확도를 높이기 위한 부하 패턴 기반의 특징 벡터 추출에 대한 연구[4]와 기존에 존재하는 고객의 전력 소비 데이터를 바탕으로 미래에 고객의 부하패턴을 예측하는 기법 연구[1] 그리고, 기존 고객의 과거 부하 프로파일을 바탕으로 시간 연관 규칙을 이용한 전력 사용 추이 분석, 주기성 탐색 연구[5] 등이 이루어지고 있다.

이 논문의 연구 내용은 다음과 같다.

첫째, 기존의 한국전력에서 분류한 계약종별에 대해 더 정확한 고객 분류와 상세한 고압 고객들의 대표 부하 프로파일 생성을 위해 통계적 군집화 기법인 기대치-최대화 알고리즘(Expectation-Maximization algorithm)을 이용한 군집화를 통하여 적절한 대표 프로파일을 결정한다.

둘째, [4]에서 소개한 3가지의 전력 부하 패턴의 특징 벡터 외에 추가적으로 국내 고압 고객의 특성을 고려한 6가지의 특징벡터를 제안한다.

셋째, 서로 다른 그룹(프로파일)에서 특정 그룹에서만 발견되는 빈발한 특징 벡터들의 패턴 발견을 위해 출현 패턴 마이닝을 적용하여, 전력 부하 패턴의 분류 모델을 생성한다.

마지막으로 한국 전력연구원에서 제공한 2007년 여름철(6월, 7월, 8월)의 전국 29개 계약종별에 대한 총 3,895개 고압 전력 사용 고객 데이터에 대해 부하 패턴 예측을 수행한다.

논문의 구성은 다음과 같다. 2장에서는 누락 데이터 및 이상치 데이터 처리와 부하 패턴의 특징 벡터 추출과 관련된 데이터 전처리 부분을 기술하고, 제 3장에서는 대표

1) 이 논문은 건설교통부 첨단도시기술개발사업 - 지능형국토정보기술혁신사업과제의 연구비지원(과제번호:07국토정보C05)에 의해 수행되었으며, 교육과학기술부와 한국산업기술재단의 지역혁신인력양성사업으로 수행하였음.

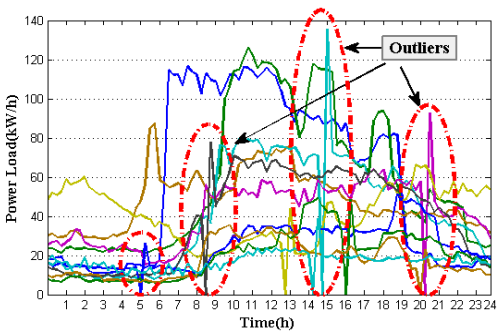
부하프로파일 생성을 위한 EM 알고리즘 및 군집 수 결정에 대해 기술한다. 출현 패턴 마이닝을 이용한 부하 패턴의 분류 기법은 4장에 기술한다. 5장에서 실험 평가를 설명하고, 6장에서 결론을 맺는다.

2. 데이터 전처리

이 논문에서 사용된 부하데이터는 한국 전력 연구원에서 구축한 배전 고압 계통 부하 분석 데이터 웨어하우스에서 전국 고압고객 144,314명을 대상으로 2007년 6월부터 8월까지 원격 검침 시스템(AMR)에 의하여 수집된 부하 데이터를 대상으로 한다. 원격 검침 시스템은 고압 고객의 전력 설비에 설치된 전자식 전력량계를 이용하여 15분 단위로 계측된 고객의 소비 전력량을 무선으로 중앙 서버에 전송한다. 정확한 마이닝 작업을 위해서 수집된 고압 고객의 부하 패턴 데이터에서 누락 데이터와 이상치 데이터를 포함하는 데이터를 제거한다. 이상치 데이터를 처리하기 위하여 정제 기법 중, SOMs[6] 군집화 기법을 적용하여 이상치 데이터를 탐지한다. SOMs 군집화 분석은 코호넨 네트워크 모델을 적용하였고, 구성 매트릭스는 10 by 10 (k=100)이다. 이상치 탐색을 위한 군집화의 입력 데이터는 식(1)과 같이 00시 부터 24시 까지 15분 간격의 총 96개의 데이터를 연결하여 벡터화 한다.

$$V_{(m,d)}^{(c)} = V_{0015}^{(c)}, \ldots, V_{h}^{(c)}, \ldots, V_{2400\,(m,d)}^{(c)} \qquad \text{식(1)}$$

(c:고객ID, h:0015,...,2400, m:월, d:일)



(그림 2) 이상치 탐색

EM 군집화 알고리즘을 사용하기 위해서 각 고객의 부하 패턴을 평일, 주말, 특수일로 분류하여 한 달 동안의 평균값을 취함으로써 고객 별로 그 달의 평일, 주말, 국경일에 대한 대표 패턴을 생성한다. 그 다음, <표 1>에서와 같이 전력 소비 형태에 대한 요약 정보를 보여줄 수 있는 부하 형태 특성(features of load shapes)[4]을 추출한다.

국내의 전력 소비 형태에 맞는 특성 벡터 추출을 위하여 <표 1>에서 P1, P2, P3, P4와 같이 배전 계통의 전문가에 의해 정의된 특정 시간대의 특징벡터[7]를 추출하였다.

<표 1> 부하 특징 벡터

특성값	정의
L1: Load Factor (24h)	$s_1 = \frac{Pattern_{Avg.for\_day}}{Pattern_{Max.for\_day}}$
L2: Night Impact (8h: 23pm~07am)	$s_2 = \frac{1}{3}\frac{Pattern_{Avg.for\_night}}{Pattern_{Avg.for\_day}}$
L3: Lunch Impact (3h: 12am~03pm)	$s_3 = \frac{1}{8}\frac{Pattern_{Avg.for\_lunch}}{Pattern_{Avg.for\_day}}$
AVG: Average for day	$avg = Pattern_{Avg.for\_day}$
MAX: Maximal value(24H)	$\max = Pattern_{Max.for\_day}$
P1:Midnight Impact (7h: 00am~07am)	$p_1 = \frac{7}{24}\frac{Pattern_{Avg.for\_night}}{Pattern_{Avg.for\_day}}$
P2: Morning Impact (3h: 09pm~12pm)	$p_2 = \frac{1}{8}\frac{Pattern_{Avg.for\_morning}}{Pattern_{Avg.for\_day}}$
P3: Afternoon Impact (3h: 13pm~17pm)	$p_3 = \frac{1}{8}\frac{Pattern_{Avg.for\_afternoon}}{Pattern_{Avg.for\_day}}$
P4: Evening Impact (4h: 19pm~23pm)	$p_4 = \frac{1}{6}\frac{Pattern_{Avg.for\_evening}}{Pattern_{Avg.for\_day}}$

3. 전력 부하 데이터의 군집화

3.1. 기대치 최대화 알고리즘

기대치 최대화(Expectation Maximization) 알고리즘은 관측되지 않은 변수의 확률모델에서 반복 연산을 통해 파라미터가 최대 우도를 갖는 값을 찾아내는 통계적 기법이다[8]. 기대치 최대화 알고리즘은 (그림 3)과 같이 2가지 단계로 구분된다.

기대치 최대화 알고리즘

1. 모델 매개변수의 초기 집합을 선택한다.
2. **repeat**
3. **기대치단계** : 각 객체가 각 분포에 속할 확률 계산한다.
4. **최대화단계** : 전 단계에서 얻은 확률이 주어지면, 기대 우도를 최대화하는 매개 변수의 새로운 추정치를 찾는다.
5. **until** 매개변수들이 변하지 않을 때 또는 매개변수의 변화가 특정 기준치 이하일 때

(그림 3) 기대치 최대화 알고리즘

주어진 $P(w_i)$를 혼합 가중치, Θ_i를 확률밀도 함수의 파라미터 성분, k를 군집의 수, X가 표본 데이터의 집합이라고 할 때, 기대치 단계에서는 식(2)와 같이 각 객체가 각 군집(분포)에 분포할 확률을 계산한다.

$$p(x|\Theta) = \sum_{i=1}^{K} p(x|\omega_i, \theta_i) P(\omega_i) \qquad \text{식(2)}$$

주어진 객체들이 서로 독립적으로 생성되었기 때문에

전 객체의 확률은 각 개체의 확률의 곱이 된다.

$$p(X|\Theta)=\prod_{i=1}^{m}\sum_{j=1}^{K}p(x_i|\omega_j,\theta_j)P(\omega_j) \qquad 식(3)$$

최대화 단계는 주어진 데이터의 확률이 가장 높은 매개변수 값을 택하기 위해 최대 우도 추정법(maximum likelihood estimation, MLE)[8]을 사용한다. MLE에 의하여 분포의 모든 매개변수와 가중치 매개 변수의 로그 우도(E)가 최대화되는 값을 찾아 클래스를 할당한다.

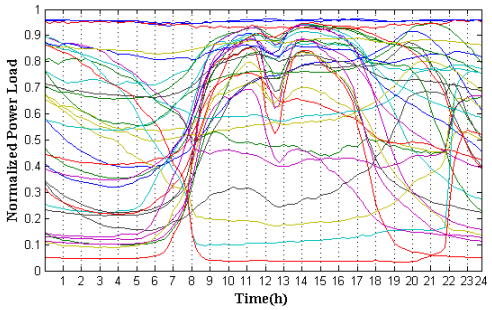
$$E=-\log L(\Theta)=-\sum_{i=1}^{m}\log p(x_i|\Theta) \qquad 식(4)$$

3.2. 교차 검증법(cross validation)을 이용한 군집수 탐색

적절한 군집수를 찾기 위하여 EM method에 우도 기반의 v-fold cross-validation 방법을 적용한다. 군집수를 결정과정은 다음과 같다.

1. 클러스터 개수를 1로 설정
2. 전체 데이터를 임의 10 등분으로 분할
3. CV 방법에 따라 10등분된 데이터를 EM method로 10번 수행
4. 10개의 결과에 대한 로그 우도의 평균 산출
5. 만약 로그 우도가 증가한다면 군집수를 1 증가 시킨 후 step 2. 을 수행

우리는 이러한 군집화 과정을 통하여 (그림4)와 같은 대표 프로파일을 생성하였다.



(그림 4) 대표 부하 프로파일

4. 출현 패턴에 의한 분류 기법

4.1. 속성값 이산화

출현 패턴 마이닝 수행을 위해서는 각 속성이 범주형 속성 값을 갖도록 이산화 되어야 한다. 추출된 특성들은 연속형 속성 값이기 때문에 이 절에서는 클래스를 고려하고 구간의 순도를 최대화하는 방식으로 분리점을 배치하는 엔트로피 기반의 이산화[9]를 적용한다.

k는 클래스의 개수이고, m_i는 분할의 i번째 구간에 속하는 값들의 수, m_{ij}를 구간 i의 클래스 j의 값의 수라고 하면, i번째 구간의 엔트로피 E_i는 다음과 같다.

$$E_i=\sum_{i=1}^{k}p_{ij}\log_2 p_{ij} \qquad 식(5)$$

$P_{ij}=m_{ij}/m_i$는 i번째 구간에서 클래스 j의 확률이며, 분할의 전체 엔트로피 E는 각 구간의 엔트로피의 가중치 평균이 된다.

$$E=\sum_{i=1}^{n}w_i e_i \qquad 식(6)$$

여기서, m은 값의 수이고, $w_i=m_i/m$은 i번째 구간의 값들의 비율이며, n은 구간의 개수이다. 연속형 속성에 대한 분할은 두 구간이 최소 엔트로피를 가지도록 초기 값들을 이분할 하는 것이다. 또한 분리 과정은 다른 구간에 대해 반복되며, 중단 기준이 만족될 때까지 구간을 선택한다.

4.2. 출현 패턴

출현 패턴이란 성장률을 분류 기준으로 둘 이상으로 분할된 데이터 집합 사이의 명확한 변화와 차이를 보이는 속성값들의 항목집합이다[10]. 성장률(growth rate)이란 두 개의 서로 다른 클래스에 해당되는 두 집합 D_1, D_2에 대한 차별 정도를 나타낸다. 패턴 X의 D_1에 대한 D_2의 성장률(GR(X))은 다음과 같이 정의 된다.

$$GR(X)=\begin{cases}0 & if\ \sup_1(X)=0\ \&\ \sup_2(X)=0\\ \infty & if\ \sup_1(X)=0\ \&\ \sup_2(X)>0\\ \sup_2/\sup_1 & otherwise\end{cases}$$

여기서, D_1을 배경(background) 데이터 집합, D_2를 목표(target) 데이터 집합이라고 하며, 출현 패턴은 배경 데이터로부터 목표 데이터 집합에 대해 높은 성장률을 가지는 패턴을 의미한다. 또한 성장률 임계값 ρ>1에 대해서 패턴 X가 GrowthRate(X)≫ρ의 성장률을 가질 때, 패턴 X를 ρ-Emerging Pattern(ρ - EP)라 한다.

출현 패턴 X의 강도(strength)는 식(7)과 같다.

$$strength(X)=\frac{GR(X)}{GR(X)+1}\times\sup(X) \qquad 식(7)$$

만약 한쪽 데이터 집합에만 나타나는 출현 패턴일 경우, 성장률(GR)이 무한대(∞)를 갖게 된다. 이러한 출현 패턴을 점핑 출현 패턴(Jumping Emerging Pattern)이라고 한다. 따라서 점핑 출현 패턴은 strength(X)=sup(X) 인 특별한 형태의 출현 패턴이다.

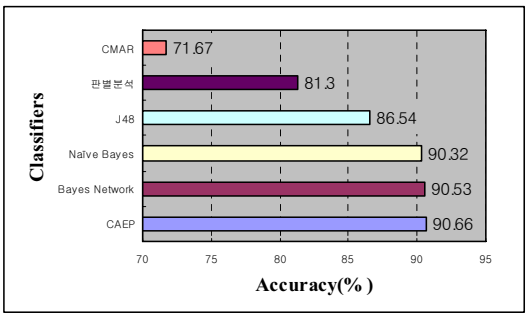
4.3. 출현 패턴에 의한 분류

모든 필수 출현 패턴 생성 후, 새로운 데이터에 대한 분류는 식(8)의 score를 계산하여 가장 높은 score 값을 가지는 클래스로 분류하게 된다[10]. 여기서, s는 분류될 데이터 인스턴스이고, E(C)는 클래스 C에서 발견된 필수 출현 패턴이다. 각각의 데이터 집합의 score를 계산하여 높은 점수를 얻는 클래스를 할당한다.

$$score(s,C) = \sum_{e \subseteq s, e \in E(C)} \sup_c(e) \times \frac{GR(e)}{GR(e)+1} \qquad 식(8)$$

5. 실험 및 평가

효율적인 군집화를 위하여 Load Factor, Night Impact, Lunch Impact와 국내의 주요 전력 소비시간 구간에 대한 4개의 특징 벡터, 그리고 전력 소비 고객의 하루 동안의 최대 소비량과 전체 평균 소비량을 추출하였다. 군집화 단계에서 추출된 특징 벡터를 바탕으로 EM 군집화 알고리즘을 적용하여 총 34개의 프로파일을 찾았으며, 각 프로파일에 대한 대표 프로파일을 만들었다. 특징 벡터의 이산화를 통하여 출현 패턴을 추출하였다. 이산화된 특징 벡터를 출현 패턴 분류기법을 적용하여 약 91%의 분류 성능을 보였다.



(그림 5) 분류 성능 비교

출현 패턴 기반의 분류기의 성능 비교를 위해 적용된 분류 기법은 의사결정트리로 C4.5, 베이지안 분류기로는 나이브 베이지안, 베이지안 네트워크 알고리즘을 적용하였다. 연관적 분류 기법으로는 CMAR[11], 통계적인 분류 기법으로는 판별분석을 적용하였다. 기타 분류 기법에 적용한 결과 Support Vector Machine이나 K-NN 기법은 공간 및 시간 복잡도 문제로 적용되지 않았다. 분류 기법의 성능을 비교한 결과 출현 패턴 기반의 분류기가 가장 높은 성능을 보였다.

6. 결론

이 논문에서는 고객의 전력 소비 패턴 분석 및 정확한 예측을 위해 고객의 전력 사용에 대한 패턴의 요약 특성을 찾았다. 그리고, 이를 통한 상당히 우수한 예측력을 갖는 주요 마이닝 작업을 제안하였다. 추출된 특성 벡터는 효율적인 군집화와 분류에 적용되었으며, 고객에 대한 주요 시간대 별 전력 소비 패턴의 형태를 찾을 수 있었다. 또, 출현 패턴 마이닝을 통하여 우리는 각각의 그룹(프로파일)에 대한 출현 패턴을 찾을 수 있었다.

이러한 특성 벡터는 향후 전력 소비 패턴의 분석과 예측에 중요한 역할을 할 것이다.

참고문헌

[1] S. J. Huang, K. Shih, "Short-term load forecasting via ARMA model identification including non-Gaussian process considerations," IEEE Trans. Power System, Vol. 18, No. 2, 2003, pp. 673-679.

[2] G. Chicco, R. Napoli, P. Postulache, M. Scutariu, C. Toader, "Customer characterization options for improving the tariff offer," IEEE Trans. Power System, Vol. 18, 2003, pp.381-387.

[3] Pitt B. and Kirchen D., 1999. "Applications of data mining techniques to load profiling", . In Proc. IEEE PICA, pp. 131-136.

[4] Figueiredo, V., Rodrigures, F., Vale, Z., Gouveia, J. B., "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques," In: IEEE Transactions on Power Systems, Vol.20, No.2, 2005, pp.596-602.

[5] Heon Gyu Lee,Bum Ju Lee,Shin Jin Ho,Long Jin,Cheng Hao Jin,Keun Ho Ryu, "Application of Calenda r-Based Temporal Classification to Forecast Customer Load Patterns from Load Demand Data", 2008 IEEE 8th International Conference on Computer and Information Technology, 2008, p. 149 ~ 154

[6] J.C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", Microsoft Research Technical Report MSR-TR-98-14, 1998.

[7] Heon Gyu Lee,Shin Jin Ho,Keun Ho Ryu, "Forecasting Electric Power Load Patterns using Unsupervised and Supervised Methods from Load Demand Data", The First International Worksop on Frontiers of Information Technology, Applications and Tools, Vol.1, 2008, p. 2-6

[8] P. S. Bradley, U. Fayyad, C. Reina, "Scaling EM (Expectation-Maximization) Clustering to Large Databases", Microsoft Research Technical Report MSR-TR-98-35, 1998.

[9] U. Fayyad, K. Irani, "Multi-Interval discretization of continuous-valued attributes for classification learning," Proc. Int'l Joint Conference. on AI, pp. 1022 - 1027, 1993.

[10] G. Dong, X. Zhang, L. Wong, J. Li, "Classification by aggregating emerging patterns," Proc. 2nd Int'l Conf. on Discovery Science, pp. 30-42, 1999.

[11] W. Li, J. Han, J. Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rule", Proceedins of the ICDM2001, pp. 369-376, 2001.