전력 부하 패턴 분석을 위한 3차원 큐브 마이닝과 캘랜더 패턴 기반 시간 데이터 마이닝

박진형*, 신진호**, Minghao Piao*, 이헌규*, 류근호*1)
*충북대학교 데이터베이스/바이오인포매틱스 연구실
**한국전력연구원 전력 정보 기술 그룹
e-mail: *{neozean, bluemhp, hglee, khryu}@dblab.chungbuk.ac.kr
**jinho@kepri.re.kr

3D Cube Mining and Calendar Pattern Based Temporal Mining for Analyzing Power Load Pattern

Jin Hyoung Park*, Jin-Ho Shin**, Minghao Piao*, Heon Gyu Lee*, and Keun Ho Ryu*
*Database/Bioinformactics Laboratory, Chungbuk National University
**Power Information Technology Group, Korea Electric Power Research Institute

요 약

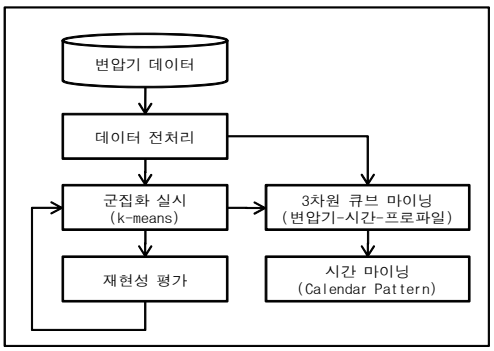
최근 전력산업에서의 에너지 가격 및 공급과 수요의 변동, 그리고 기후의 변화에 의해서 부하 예측은 전력회사 경영방침 계획에 있어 중요한 요소가 되었다. 이 논문에서 전력계통의 최적 운용 계획을 위하여 우리가 제안한 기법은 다차원 분석이 가능한 3D 큐브 마이닝과 시간의 변화에 따른 패턴 예측이 가능한 캘린더 기반 시간 데이터 마이닝 기법이다. 이를 통하여 무선 부하 감시 시스템의 부하 데이터의 다차원 분석이 가능하고, 시간 변화에 따른 서로 다른 부하 패턴의 예측이 가능하도록 한다.

1. 서론

전력 부하 예측을 위한 정확한 분석 모델은 전력 운영과 계획에 필수적이며 전력산업에 있어서의 부하 예측은 전력의 구입, 생산, 기반 시설의 업데이트 등에 대한 의사결정을 함에 있어서 도움을 준다. 특히, 부하 예측은 공급자, ISO 회사 및 재무기관이 전기에너지 생산, 전송, 분배, 및 관련 마케팅에 있어서 매우 중요하다. 따라서 전력산업에 있어서 효율적인 운용과 계획을 위해 정확한 부하 및 부하 패턴 예측 기술이 필요하며, 이를 위해 통계[1] 및 데이터마이닝[2, 3] 등과 같은 수학적 방법들이 부하 분석 모델링을 위해 사용된다. 전력시스템에서의 데이터마이닝 기술은 부하데이터로부터 규칙성을 인지하고 추출하는 가장 대표적인 기술이며 데이터마이닝 기법 적용 결과인 패턴 규칙 집합은 부하 데이터로부터 이전에 알려지지 않은 부하 패턴을 식별할 수 있게 한다. 또한 부하 패턴 규칙 집합은 현재의 부하패턴과 분석을 위해서 비교 되어 질 수 있다. 일반적으로, 데이터마이닝 기술을 적용한 부하패턴 예측은 관련된 정보로부터 부하패턴 모델을 생성하고 이 모델을 적용하여 새로운 부하패턴을 예측한다. 부하 예측의 범주로는 시간 단위나 주일(week)을 단위로 하는 단기예측, 일주일부터 일 년을 단위로 하는 중기예측, 그리고 일 년 이상을 단위로 하는 장기예측 등 세 가지로 나뉜다.

이 논문에서는 무선 감시 시스템으로부터 30분 간격으로 측정된 변압기 부하 데이터의 부하 프로파일 생성과 부하패턴의 예측을 위해, 기존의 데이터마이닝 기법에 시간적 의미와 관계를 표현할 수 있고, 시간의 변화에 따른 서로 다른 부하 패턴 분석이 가능한 시간 데이터마이닝 기법적용을 제안한다. 또한 캘린더 패턴 기반의 시간 마이닝 적용으로 "년, 월, 주, 일"과 같은 사용자 기반의 시간 단위 제약조건 설정을 통해 단기, 중기, 장기 기간의 부하 패턴 발견이 가능하다. 제안한 변압기 부하 패턴 분석을 위한 시간 데이터 마이닝 기법 적용의 단계는 그림 1과 같고 그 세부 내용은 다음과 같다.

1) 변압기 대표 부하 프로파일을 위한 군집화 및 재현성 평가
2) 3차원 「시간-변압기-프로파일」 패턴 발견을 위한 3D 데이터 큐브 마이닝 적용
3) 시간적 특성을 고려한 변압기 부하 패턴 분석 및 향후 부하 패턴 예측을 위한 캘린더 패턴식의 시간 데이터마이닝 알고리즘의 적용



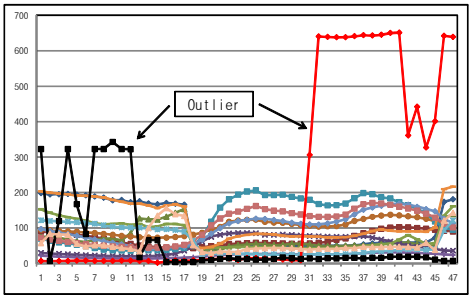
(그림 1) 부하 패턴 분석 단계

1) 이 논문은 전력연구원 부하분석모델 시공간 데이터마이닝 기법적용연구 과제와 2008년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (R01-2007-000-10926-0).

2. 부하 데이터 수집 및 데이터 전처리

부하 데이터는 기 구축된 GIS-AMR 기반의 배전 고압계통 부하분석모델의 부하 분석 데이터웨어하우스[4]에서 서울 강남구 지역에 설치된 총 45,884개의 변압기중 무선 부하감시센서가 부착된 5,071개의 변압기의 2007년 1월부터 10월까지의 30분 간격으로 측정한 변압기 부하량를 분석 대상으로 한다.

30분 간격의 계측된 부하량은 미계측값 또는 데이터 입력시의 오류 때문에 원시 데이터에는 많은 이상치 (그림 2)를 포함하고 있다. 이러한 이상치 데이터는 마이닝 수행결과에 대한 신뢰성을 보장하기 어렵기 때문에 데이터 정제를 위한 전처리 과정을 거쳐야 한다.



(그림 2) 이상치 부하량 데이터

수집된 변압기 부하 패턴 데이터 집합에 포함된 이상치 데이터 처리를 위하여 정제 기법 중, *SOMs*[2] 군집화 기법을 적용하여 이상치 데이터를 탐지한다. 이상치는 유사한 값들의 그룹들로 구성되는 군집화에 의해 탐지될 될 수 있으며, 군집들의 집합 밖에 위치하는 값은 이상치로 간주한다. SOMs 군집화 분석은 코호넨 네트워크 모델을 적용하였고, 구성 매트릭스는 10 by 10(100)이다. 이 중 한 클러스터에 포함된 데이터 객체가 1~3개 이하인 군집의 부하 패턴에 대해 이상치로 간주하여 제거시킨다.

3. 변압기 대표 부하 프로파일 생성

전처리된 일(day) 단위의 변압기 데이터로부터 k-means 군집분석을 수행하여 대표 부하 프로파일(profile)을 생성한다. 군집화 기술 중 대표 패턴 생성에서의 k-means 사용은 빠른 군집의 구성과 사용자 기반의 군집 수 결정, 그리고 k개의 군집의 적합성을 판단하는 재현율 적용이 용이하기 때문이다.

3.1 k-means 알고리즘을 이용한 부하 패턴 군집화

k-means[5] 알고리즘 적용은 추출된 전기 부하량을 변압기별, 일별로 구분하여 시간에 대한 부하량의 vector로 다음 식과 같이 표현된다.

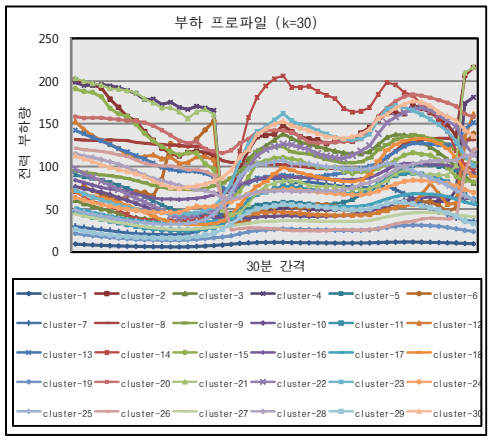
$$V^{(b^{th})} = V_0^{(b^{th})}, \ldots, V_t^{(b^{th})}, \ldots, V_T^{(b^{th})} \quad \text{(식 1)}$$

(단, b^{th} = bqjsWo, 변압기, t = 0,...,2330)

(그림 3)은 vector로 표현된 모든 부하 패턴으로부터 k-means 군집분석을 수행하여 생성된 대표 프로파일이다.

3.2 최적 군집 개수 결정을 위한 재현성 평가

변압기의 일별 부하량에 대한 군집의 적합성을 판단하기 위해서 신경망이나 결정트리 및 회귀와 같은 지도학습 모델링에서 사용되는 데이터 분할 기법[6]을 이용한다. 자료 분할은 동일한 군집화 방법의 반복을 가능하게 해주므로 이를 활용하여 재현성을 평가하며, 재현성 평가 절차알고리즘은 (그림 4)과 같다.



(그림 3) 변압기 대표 부하 패턴

1. 데이터를 임의의 2개로 분할한다. 하나를 훈련 데이터, 다른 하나를 테스트 데이터라고 가정한다.
2. 훈련 데이터를 군집화하여 모델을 산출한다. 그리고 테스트 데이터의 각 개체를 훈련 데이터로부터 생성된 군집화 모델에 적용하여 분리한다. 즉 테스트 데이터는 가장 가까운 중심의 군집에 속하게 된다.
3. 테스트 데이터를 동일한 방식으로 군집화하여 자체 모델을 산출한다. 이에 따라 테스트 데이터의 각 개체는 몇 개의 군집 중 하나로 배속된다.
4. 테스트 데이터의 군집화 결과를 토대로 교차분류표를 만든다. 적용된 군집화가 최적화된 것이라면 이 표에서 행과 열은 강한 대응성을 보일 것이다. 그러나 그렇지 않다면 행과 열의 대응성은 약하게 나타날 것이다.

(그림 4) 재현성 평가 알고리즘

재현성 평가 후 최종 군집수의 결정은 교차분류표에서 주 경향에서 벗어난 데이터의 percentage가 가장 작은 k 값을 군집수로 결정한다. 변압기 데이터의 경우 재현성 평가 결과 k에 대한 재현율은 <표1>과 같다.

<표 1> 변압기 부하 패턴재현성 평가 결과

군집 개수	25	29	**30**	31	33
재현율	22.9	24.33	**19.4**	31.32	33.99

k = 30일 때에 가장 낮은 재현율을 보이기 때문에 변압기 군집화시 k의 값을 30으로 설정한다.

4. 3D 데이터 큐브 마이닝을 이용한 「시간-변압기-프로파일」 패턴 발견

변압기별 시간에 변화에 따른 부하 패턴 분석을 위해서 3차원 큐브 마이닝 기법[7]을 적용한다. 기존의 2D FCP(frequent closed pattern) 알고리즘은 2차원 매트릭스 표현 데이터를 이용하여 빈발 항목집합을 발견하였다. 따라서 3차원 부하패턴 데이터에서 시간 속성을 고려한 빈

발 항목집합을 발견하는데 적용할 수는 없으므로 [7]에서 제안한 RSM 알고리즘을 적용하여 3차원 「시간-변압기-프로파일」 데이터의 빈발 패턴을 탐사한다. 3차원 큐브 마이닝에 변압기의 시간에 변화에 따른 부하 패턴을 적용하기 위해서 3장에서 군집화한 데이터로부터 <표 2>과 같이 변압기 ID, 일시 ID, 부하 프로파일(패턴) 속성의 기본 데이터를 구성한다.

<표 2> 변압기 부하 패턴 데이터

속성명	데이터 타입	설명
bank_id	nominal	뱅크 고유 식별 코드
date_time_id	continuous	일시 지정 코드
profile	nominal	부하 패턴 군집 번호

변압기 ID, 일시 ID, 부하 패턴 군집 번호를 3차원 속성으로 지정하고, 지정한 속성을 순차 정렬한다. 이를 인덱스 형태로 전환(<표 3>)하여 기본 데이터를 <표 4>와 같이 0과 1로 구성된 순차적인 큐브형 데이터로 변환한다.

<표 3> 각 속성의 인덱스 변환

date_time_id	bank_id	profile
36961	22	cluster-19
37057	24	cluster-27

↓

date_time_id	bank_id	profile
h1	r1	c19
h2	r2	c27

<표 4> 3차원 큐브형 데이터 구성 예제

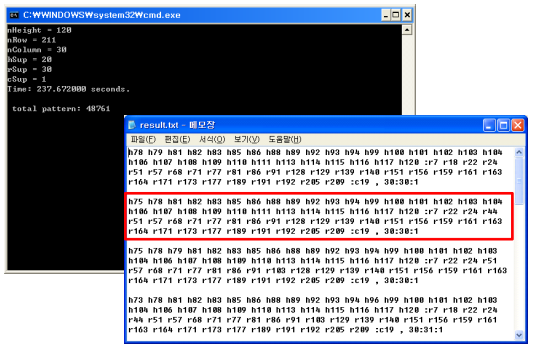
date_time_id	bank_id	profile			
		c1	c2	...	c_n
h1	r1	0	1	0	0
	r2	1	0	0	0
	...	0	1	0	0
	r_m	0	0	0	1
h2	r1	0	0	1	0
	r2	0	1	0	0
	...	0	0	1	0
	r_m	0	0	1	0
...	r1	1	0	0	0
	r2	1	0	0	0
	...	0	1	0	0
	r_m	0	0	1	0
h_k	r1	0	1	0	0
	r2	0	0	1	0
	...	0	0	0	1
	r_m	1	0	0	0

3차원 큐브 마이닝 알고리즘을 실행하기 위해서는 3가지 속성(row, column, height)에 대한 사용자 지지도 임계값을 갖는다. <표 6>는 3가지 속성의 지지도 변화에 따른 빈발 패턴이며, 이 논문에서는 11,162개인 임계값을 적정한 값으로 선택한다.

(그림4)은 큐브 마이닝 기법을 적용한 빈발 패턴 예이다. 시간 집합 *T*를 *{h75, h78, h82, ..., h120}*, 변압기 집합 *B*를 *{r7, r22, ..., r209}*라고 할 때, *cluster-19*는 변압기 집합 *B*가 기본 시간 집합 *T*에서 빈발하다고 해석된다.

<표 6> 각 속성에 대한 임계값과 빈발 패턴 수

Height	Row	Column	패턴수
30	40	1	16
35	35	1	115
30	30	1	11162
20	30	1	48761
20	20	1	354752



(그림 4) 큐브 마이닝 수행 결과 예제

5. 시간의 변화에 따른 부하 패턴 분석을 위한 캘린더 기반 시간 마이닝

3차원 큐브 마이닝 적용 결과에서 기본 시간 간격에서의 패턴에 대해 달력 표현과 주기성을 표현하기 위해서 기본 시간을 캘린더 스키마에 기반한 시간 패턴으로 갱신한다. 캘린더 스키마는 달력의 개념 계층에 의해 결정되어지고 유효성 제약조건을 갖는 관계형 스키마이다[8].

캘린더 스키마(*CS : Calendar Schema*) : 캘린더 스키마는 달력 표현의 시간 단위와 그 단위에서의 가능한 도메인의 집합으로 정의되며, 그 형태는 다음과 같다.

$$CS = (G_n : D_n, G_{n-1} : D_{n-1}, ..., G_1 : D_1) \quad \text{(식 2)}$$

$1 \le i \le n$에 대해, 속성 G_i는 년, 월, 일 등과 같은 달력 개념에서의 시간 단위이고, 각 D_i는 양의 정수의 유한 집합으로 속성 G_i의 도메인 값의 집합을 나타낸다.

캘린더 패턴(*CP: Calendar Pattern*) : 캘린더 패턴은 주어진 스키마 $CS = (G_n : D_n, ..., G_1 : D_1)$의 인스턴스이며, $CP = \{d_n, ..., d_1\}$으로 표현된다. 여기서 각 d_i는 D_i의 도메인 값이거나 문자 '*'이다. 만약 d_i가 '*'이라면 그 의미는 도메인 D_i의 모든 값을 나타내고 "every"로 해석한다. 또한 캘린더 패턴에 i개의 '*'를 가지는 패턴은 *i-star* 패턴이라 부르고 그 집합을 $\Phi(CP_i)$로 나타낸다. 특히 '*'를 전혀 포함하지 않는 캘린더 패턴에 대해서는 "기본시간단위"라고 부른다.

캘린더 패턴 갱신(*CP: Calendar Pattern Update*) : 변압기-프로파일에 대한 빈발 패턴(*FP : Frequent Pattern*)를 시간에 대한 패턴이라고 할 때, i-star 패턴으로의 갱신 알고리즘은 (그림 5)와 같다. 각 i-star 패턴에 대한 규칙들은 각각의 해당되는 카운터를 가지고 있다. 만약 갱신 과정이 처음 이루어지는 단계(3~5라인)이거나 기본시간단위에서의 규칙이 처음 갱신되어질 경우(7~9라인), $FP_k(CP_0)$를 $FP_k(CP_i)$으로 할당하고 카운터를 1로 설정한다. 그렇지

않은 경우(10~12라인), 현재의 카운터를 1씩 증가시킨다. 만약 CP_i에 의해 포함되는 CP_0가 N_i개, 현재 $FP_k(CP_i)$으로 n번째 갱신 단계라고 가정할 경우, $FP_k(CP_i)$의 각 카운터는 공식, $updateCnt + (N_i\text{-}n) \geq \min Fre \cdot N_i$을 만족하는 $FP_k(CP_i)$들만을 유효한 규칙으로 생성한다.

```
1  input FPk(CP0),CP0, minFre
2  output FPk(CPi),CPi satisfy minimum frequency
3    if FPk(CPi) is first time updated then
4      FPk(CPi) = FPk(CP0);
5      updateCnt = 1;
6    else
7      for each FP ∈ FPk(CP0)- FPk(CPi)
8        FP.updateCnt = 1;
9      for each FP ∈ FPk(CP0)∩FPk(CPi)
10       FP.updateCnt++;
11   FPk(CPi) = {FP.updateCnt ≥ minFre·Ni}
12   end if
```

(그림 5) *i-star* 캘린더 패턴 갱신 알고리즘

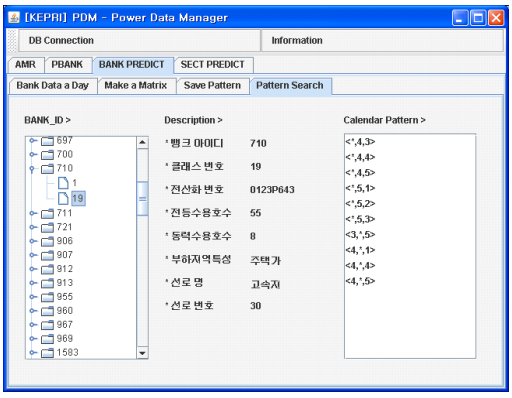
5.1 기본 시간에 대한 달력 표현

변압기 부하 패턴 예측에 사용된 캘린더 스키마는 $CS = (G_n : D_n, ..., G_1 : D_1) \Rightarrow CS = (Month : \{1\sim4\}, Week : \{1\sim4\}, Day : \{1\sim7\})$이다. 다음 <표8>에서 사용된 캘리더 스키마에 대한 의미 해석을 보여주고 있다.

<표 8> 캘린더 패턴에 대한 의미 해석

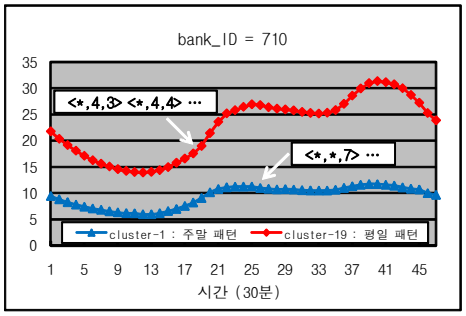
속성	Month	Week	Day
도메인	{1,2,3,4}	{1,2,3,4}	{1,2,3,4,5,6,7}
의미	1: 1월, 2: 2월, ...	1: 첫주, 2: 둘째주, ...	1: 월요일, 2: 화요일, ...

1월~10월까지의 변압기 부하 데이터에 캘린더 패턴으로의 갱신 후는 (그림 5)와 같다.



(그림 5) 캘린더 시간 패턴 갱신 결과

(그림 6)은 시간 데이터마이닝 기법을 적용한 변압기 부하 패턴 예측에 대한 예이다. 그림에서 Bank ID=710인 Bank는 '매월 매주 일요일'에 대표 패턴(cluster-1)을 갖으며, '매월 넷째 수요일 목요일'에 cluster 19번의 대표 패턴을 갖는다는 의미이다.



(그림 6) 변압기 710의 주중/주말 시간 패턴 및 부하 패턴

6. 결론

이 논문에서는 실제 전력 계통의 최적운영을 위하여 시간의 변화에 따른 전력 부하 패턴을 분석하였고, 이를 위하여 3차원 데이터 큐브 마이닝과 캘린더 패턴 기반의 시간 데이터마이닝을 적용하였다. 제안된 시간 데이터마이닝의 기법 적용은 주중 또는 주말과 같은 서로 다른 시간대의 변압기 부하 패턴 분석과 특정 시간의 주기성을 갖는 주기적 부하 패턴 분석이 가능하므로 향후 유사한 변압기 정보를 가진 변압기의 전력 부하 패턴 예측이 가능하다.

참고문헌

[1] Huang S. J., Shih K. R., "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] Verdu S.V., "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps", IEEE Trans. Power System. Vol. 21, No. 4, 2006, pp. 1672-1682

[3] Chicco G., Napoli R., Piglione F., Postolache P., Scutariu M., "Load pattern-based classification of electricity customers", IEEE Trans. Power System. Vol. 19, No. 2, 2004, pp. 1232- 1239

[4] "인터넷 GIS 환경의 AMR 시스템 연계 모델 개발", 전력산업연구개발 보고서, 전력연구원, 2006.

[5] Kanungo T., et al., "An efficient k-means clustering algorithm: analysis andimplementation", IEEE Trans. Machine Intelligence. Vol. 24, No. 7, 2002, 881-892

[6] William M. Rand, "Objective Criteria for the Evaluation of Clustering Methods", Journal of the American Statistical Association, Vol. 66, No. 336, 1971, pp. 846-850

[7] Liping Ji, Kian-Lee Tan, Anthony K. H. Tung, "Mining frequent closed cubes in 3D datasets", Proceedings of the 32nd international conference on Very large data bases, pp. 811 - 822.

[8] 이헌규, 노기용, 서성보, 류근호, "캘린더 패턴 기반의 시간 연관적 분류 기법", 정보과학회 논문지, Vol. 32, No. 6, 2005, pp. 567 - 584.