고객 공통 정보를 이용한 데이터마이닝 기반의 고객 분류 기법에 대한 연구

김영일, 송재주, 양일권
한국전력공사 전력연구원 녹색성장연구소

Study of Customer Classification Algorithm Based on Data Mining Technology Using Customer Common Information

Young-Il Kim, Jae-Ju Song, Il-Kwon Yang
Green Growth Laboratory, Korea Electric Power Research Institute, KEPCO

Abstract - 자동검침 데이터를 이용하여 고객의 가상 부하패턴을 생성하고 회선 및 구간의 부하를 분석하는 연구가 활발히 진행되고 있다. 본 논문에서는 기존에 연구된 산업분류 별 평균 부하패턴을 이용하는 방법과 고객의 부하 형태 인덱스를 이용한 방법의 문제점을 살펴보고, 이를 개선하기 위한 방법으로 고객의 속성정보를 이용하여 고객을 분류하는 방법을 제안하였다.

1. 서 론

초기의 자동검침시스템은 검침 비용 절감을 위해 도입되었으나 최근에는 배전선로의 회선 및 구간에 대한 부하정보 분석에 많이 활용되고 있다. 기존에는 SOMAS (Substation Operating results MAnagement System)를 통해 회선에 대한 15분 단위의 부하분석만 가능하였으나, 최근에는 15분 단위의 자동검침을 수행한 고객들의 데이터를 이용하여 구간 단위의 부하를 분석하는 연구가 활발히 진행되고 있다 [1]-[4]. 구간의 부하는 구간에 연결된 자동검침 고객과 미 검침 고객의 부하의 합으로 이루어진다. 자동검침 고객의 데이터를 군집화하고 각 군집 별로 대표 부하패턴을 만들어, 미 검침 고객을 데이터마이닝의 분류기법을 이용하여 각 군집들 중에 하나의 군집에 할당되도록 분류하여 해당 군집의 대표 부하패턴에 미 검침 고객의 월 사용량을 적용하면 가상 부하패턴을 만들 수 있게 된다. 따라서 구간의 가상 부하패턴의 정확도를 높이기 위해서는 미 검침 고객의 가상 부하패턴의 정확도를 높이는 것이 중요하다. 본 연구에서는 기존에 연구된 고객 분류방식을 살펴보고, 각 방식들의 문제점을 살펴보고, 이를 개선하기 위한 방법을 제안하였다.

2. 본 론

2.1 관련연구 분석

2.1.1 군집화 및 분류의 기본 방법

미 검침 고객의 가상 부하패턴을 생성하기 위해서는 다음과 같은 4가지 단계를 거치게 된다.

① 자동검침 고객의 군집화를 통한 대표 부하패턴 생성
② 모든 고객이 공통으로 갖는 정보를 이용하여 하나의 군집에 매칭할 수 있는 분류 알고리즘 생성
③ 미 검침 고객의 공통 정보를 입력으로 하여 분류 알고리즘을 통해 대표 부하패턴 선택
④ 대표 부하패턴에 미 검침 고객의 월 전력사용량을 대입하여 가상 부하패턴 생성

2.1.2 기존의 고객 분류 방법

연구논문 [5]에서는 산업분류 별 평균 부하패턴을 이용한 분류 방법을 제안하고 있다. 군집화를 위한 알고리즘은 Fuzzy C-Means (FCM) 알고리즘과 Hierarchical Clustering (HC) 알고리즘을 비교하여 대표 부하패턴의 정확도가 높은 FCM을 선정하였다. 고객의 공통 정보는 고객이 속한 산업분류의 평균 부하패턴(ALP: Average Load Profile)을 이용하였으며 Probability neural networks (PNN)을 적용하여 고객을 분류하였다. 이 방법은 상이한 부하패턴을 갖는 고객이 동일한 산업분류 코드를 갖는 경우 대표 부하패턴의 오차가 크게 발생하는 문제점을 갖게 된다.

연구논문 [6]에서는 부하형태 인덱스(LSI: Load Shape Index)를 이용한 분류 방법을 제안하고 있다. 군집화를 위한 알고리즘으로 K-Means 알고리즘을 사용하고, 고객의 공통 정보는 LSI를 이용하였으며 C5.0 알고리즘을 적용하여 고객을 분류하였다. 이러한 방법은 자동검침 고객의 경우에는 검침 데이터를 이용하여 LSI를 계산할 수 있으나, 미 검침 고객의 경우에는 LSI를 계산하기 위한 데이터베이스가 구축되어 있지 않은 경우에는 적용할 수 없는 문제점이 있다.

2.2 고객 속성정보를 이용한 분류 방법

본 논문에서는 기존의 미 검침 고객의 분류방법에 대한 문제점을 해결하기 위한 방법으로 고객의 다양한 속성정보를 이용한 분류 방법을 제시하고자 한다. 앞에서 살펴보았듯이 산업코드를 이용한 고객분류 방법은 동일한 산업분류에 속한 고객이라고 하더라도 부하 곡선이 상이한 경우가 많이 있으므로 고객의 산업코드 하나만을 이용하여 고객을 분류하는 방식은 분류과정에서 많은 오차를 갖게 된다. 분류를 위해 LSI를 이용하는 방법은 고객의 점심 시간대와 저녁 시간대의 부하 특성을 이용하여 분류하게 되므로 산업코드를 이용한 방법에 비해 높은 분류 성능을 보이게 된다. 그러나 자동 검침 고객의 경우에는 검침값을 이용하여 LSI를 계산할 수 있으나, 미 검침 고객의 경우에는 LSI의 값을 알 수가 없기 때문에 고객 조사 등을 통해 각 고객의 부하 특성을 조사하여 데이터베이스를 구축해야만 적용이 가능한 단점을 갖게 된다.

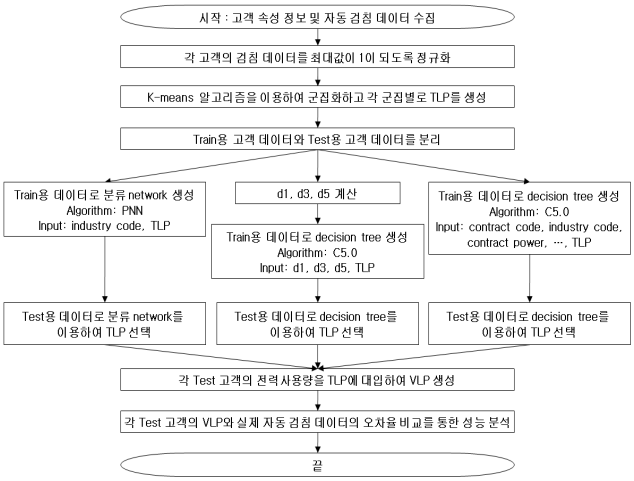
<표 1> 정규화된 부하형태 인덱스 (LSI)

Parameter	Definition	Period of definition
Load Factor	$d_1 = \frac{P_{av,day}}{P_{max,day}}$	1 day
Night Impact	$d_3 = \frac{1}{3}\frac{P_{av,night}}{P_{av,day}}$	1 day (8 hours from 23:00 to 07:00)
Lunch Impact	$d_5 = \frac{P_{av,lunch}}{P_{av,day}}$	1 day (3 hours from 12:00 to 15:00)

본 논문에서는 자동검침 고객의 부하패턴을 데이터마이닝의 군집기법을 이용하여 분류된 각 군집에 미 검침 고객을 매칭하기 위한 고객 공통 정보로 이미 전력회사의 고객 데이터베이스에 구축된 속성 정보들 중에서 다음과 같은 정보를 선택하여 고객을 분류하기 위한 입력 값으로 사용하였다.

- 계약종류 코드 (주택용, 일반용, 교육용, …)
- 전기사용 용도 (아파트, 가로등, 사무실, …)
- 계약전력량 (10kW, 30kW, 100kW, …)
- 산업분류 코드 (160여 개)
- 전기공급방식 (단상2선, 단상3선, 삼상3선, …)

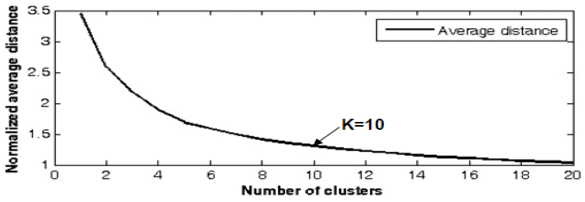
본 논문에서는 앞에서 소개한 두 가지 방법과 본 논문에서 제안한 방법에 대한 성능 비교를 위해 그림 1과 같은 방법으로 실험하였다.



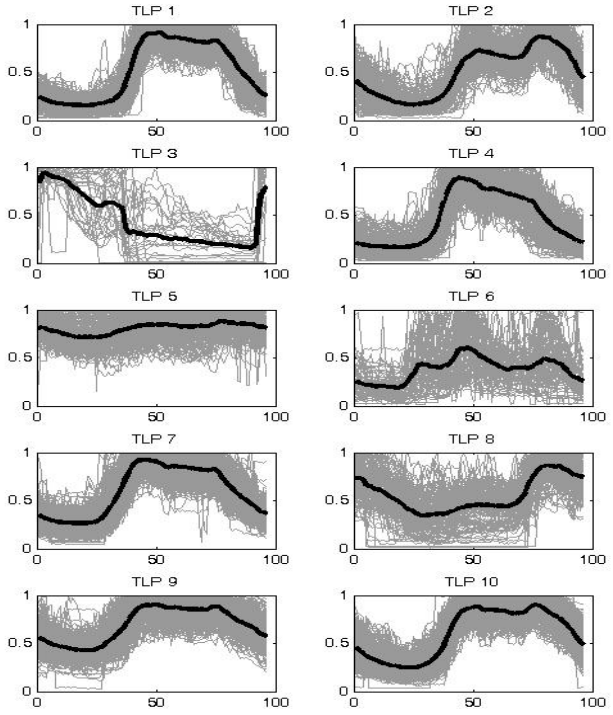
<그림 1> 세 가지 고객 분류 방법의 성능 비교를 위한 실험 절차

2.3 실험 결과 분석

본 논문에서는 실험을 위해 강남지점의 2878명의 자동 검침 고객을 대상으로 데이터를 수집하였다. 고객의 군집화를 위해 k-means 알고리즘을 사용하였으며, 군집의 수인 k의 값은 각 고객과 고객이 속한 군집의 중심 노드와의 평균 오차를 계산하여 이 값이 수렴하는 지점인 10을 선정하였다. K-means 알고리즘을 통해 생성된 10개의 군집의 TLP는 그림 3과 같다. 그림에서 TLP는 굵은 선으로 표시되었으며, 회색 실선은 각 군집에 속한 고객의 부하곡선을 나타낸다.



<그림 2> k값 변화에 따른 평균 오차 그래프



<그림 3> K-means 알고리즘으로 군집화된 10개의 TLP 그래프

군집화된 결과를 분류 기법을 이용하여 고객을 분류하기 위해 전체 데이터의 10%를 테스트용 데이터로 구분하였다. 표 2는 세 가지 고객 분류 방법을 이용하여 훈련용 데이터로 10개의 군집에 대한 고객 분류 기준을 만들고, 테스트용 데이터로 시뮬레이션 하였을 경우에 정확하게 예측할 확률을 나타낸다. 다양한 고객 속성 정보를 이용하는 방식이 산업코드만을 이용하는 방식에 비해 좀 더 많은 분류 기준을 제시할 수 있어서 고객 분류에 있어서 좀 더 나은 성능을 보이는 것을 알 수 있다. LSI 방식은 다른 두 방식에 비해 월등히 높은 성능을 보이고 있으나, 이는 고객의 부하 곡선의 형태를 미리 알고 있어야 한다는 단점을 갖고 있어 시스템적으로 운영하기에는 어려운 점이 있다. 분류 방식의 정확도를 좀 더 높이기 위해서 TLP의 형태가 유사한 1, 7 군집과 9,10 군집을 각각 하나의 군집으로 묶어 8개의 군집으로 만들어 분류하도록 실험해 보았다. 8개의 군집으로 분류할 경우의 정확도는 표 2에서 보는 바와 같이 유사한 형태의 군집을 하나로 합침으로써 분류에 따른 오차가 좀 더 줄어든 것을 알 수 있다.

<표 2> 군집 수에 따른 분류 정확도

분류 방법	10개 군집의 경우 분류 정확도	8개 군집의 경우 분류 정확도
산업코드 방법	0.18	0.35
LSI 방법	0.51	0.63
제안된 방법	0.26	0.42

표 3은 10개의 군집과 8개의 군집을 사용했을 경우의 테스트용 고객의 실제 부하와 계산을 통해 얻은 VLP의 평균 오차를 나타낸다. 실험 결과를 살펴보면 군집이 10개인 경우와 8개인 경우의 평균오차의 차이가 별로 없음을 알 수 있으며, 이는 분류 과정에서 발생한 정확도는 많이 차이를 보이지만 그 차이는 고객이 속한 군집과 유사한 군집으로 분류되어 발생한 오차이기 때문에 평균 오차의 변화는 미묘한 것임을 알 수 있다.

<표 3> 군집 수에 따른 실제 부하와 가상 부하패턴 간의 평균 오차

분류 방법	10개 군집의 경우 평균 오차	8개 군집의 경우 평균 오차
산업코드 방법	0.196	0.193
LSI 방법	0.143	0.142
제안된 방법	0.178	0.176

세 가지 분류 방식의 평균 오차를 살펴보면, 산업코드를 이용한 방식이 가장 오차가 크고, 본 논문에서 제안된 방식이 산업코드를 이용하는 방식보다는 오차가 적은 것을 알 수 있다. LSI 방식의 경우에는 본 논문에서 제안된 방식보다 오차는 적지만 고객의 부하 형태에 대한 인덱스 값을 미리 알고 있어야 적용이 가능하므로, 전력사에서는 본 논문에서 제안된 방식을 이용하여 별도의 고객에 대한 조사 없이 기존에 구축된 고객 속성 정보만을 이용하여 효과적으로 미 검침 고객에 대한 가상 부하패턴을 계산할 수 있게 된다.

3. 결 론

미 검침 고객에 대한 가상 부하패턴을 계산하는 방법은 배전선로의 부하 예측에 있어서 중요한 역할을 한다. 본 논문에서는 기존에 제안된 고객 분류 방식의 문제점을 살펴보고, 이를 해결하기 위한 방법으로 고객의 속성 정보들을 이용한 분류 방법을 제안하였다. 본 논문에서 언급된 세 가지 고객 분류 방법을 동일한 데이터를 이용하여 적용하고 각각의 분류 방법의 분류 정확도와 분류된 고객의 실제 부하곡선과 가상 부하곡선의 평균 오차를 계산하여 성능을 평가하였다. 본 논문에서 제안된 방법은 산업코드를 이용하는 방법에 비해 정확도나 평균 오차의 관점에서 나은 성능을 보였다. 고객의 부하 형태를 나타내는 LSI를 이용하는 방법에 비해서는 성능에 낮았으나, LSI를 이용하는 방법은 미 검침 고객에 대한 LSI를 미리 알고 있다는 가정을 만족시켜야 하므로, 전력사가 고객 조사를 통해 LSI를 데이터베이스로 구축해야 하므로 실제 시스템에 적용하기에는 비용적인 부담이 크게 된다. 따라서 본 논문에서 제안된 방법을 통해 기존의 고객 속성 정보를 활용하여 보다 정확한 가상 부하패턴을 계산할 수 있어 배전선로의 부하 예측에 많은 기여를 할 것으로 생각된다.

[참 고 문 헌]

[1] Koo-Hyung Chung, Chan-Joo Lee, Jin-Ho Kim, Don Hur, Balho H. Kim, and Jong-Bae Park, "Development of Customer Oriented Load Management Software for Saving on Utility Bills in the Electricity Market", Journal of Electrical Engineering & Technology, 2007, Vol. 2, No. 1, pp. 42-49.

[2] Chongqing Kang, Xu Cheng, Qing Xia, Yonghao Huang, and Feng Gao, "Novel approach considering load-relative factors in short-term load forecasting", Electric Power Systems Research, Vol. 70, Issue 2, July 2004, pp. 99-107.

[3] 신진호, 김영일, 송재주, 이봉재, 이정일, "지리정보와 검침데이터를 이용한 배전계통 부하분석모델 개발", 대한전기학회 하계학술대회, 2006, 7월, pp. 2124-2125.

[4] G. W. Chang, S. Y. Chu, and H. L. Wang, "A Simplified Forward and Backward Sweep Approach for Distribution System Load Flow Analysis," 2006 International Conference on Power System Technology, pp. 1-5.

[5] David Gerbec, Samo Gasperic, Ivan Smon, and Ferdinand Gubina, "Allocation of the Load Profiles to Consumers Using Probabilistic Neural Networks", IEEE Transactions on Power Systems, Vol. 20, No. 2, May 2005, pp. 548-555.

[6] Vera Figueiredo, Fatima Rodrigues, Zita Vale, and Joaquim Borges Gouveia, "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques", IEEE Transactions on Power Systems, Vol. 20, No 2, May 2005, pp. 596-602.