Smart Media Journal / Vol.8, No.4 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2019.8.4.80

AMI시스템에서 유사도를 활용한 누락데이터 보정 방법

Estimate method of missing data using Similarity in AMI system

권혁록*, 홍택은**, 김판구***
(Hyuk-Rok Kwon, Taek-Eun Hong, Pan-Koo Kim)

요 약

AMI가 확대보급이 빠르게 진행되고 있고, 이에 따라 전력사용 데이터를 활용한 다양한 서비스들이 늘어나고 있다. 이러한 서비스를 효용성을 높이기 위해서 누락된 계량데이터들을 보정할 필요가 있다. 본 논문에서는 누락된 계량데이터의 보정을 위해서 유클리디안 유사도를 이용하여 사용량 패턴이 유사한 고객을 찾아 누락데이터를 보정하는 방식을 제안하고 선행 방식과의 비교자료를 제공한다.

■ 중심어 : 추정 ; 유사도 ; AMI ; 누락데이터

Abstract

As a result of AMI rapidly expanding and distributing its products, variety of services that utilize data on the use of electricity are increasing. In order to make these services more effective, missing metric data needs to be corrected, compensating for which Euclidean similarity is used to find customers with similar usage patterns. Throughout such a process, we propose a method for correcting missing data and provide comparison with the preceding methods.

■ keywords : Estimate ; Similarity ; AMI ; MissingData

Ⅰ. 서 론

AMI(Advanced Metering Infrastructure, 지능형 계량 인프라)는 스마트그리드를 구현하기 위해 필요한 핵심 인프라로서 스마트미터, 통신망, MDMS(Meter Data Management System, 계량데이터관리시스템)와 운영시스템으로 구성되고 스마트미터 내에 모뎀을 설치하여 양방향 통신이 가능한 지능형 전력계량인프라를 말한다. AMI 운영시스템에서는 소비자와 전력회사 간 양방향통신으로 원격검침, 수요관리, 전력소비 절감과 전기품질 향상 등 다양한 융복합 서비스를 제공하게 된다.

한국전력은 에너지신산업 가속화 정책에 따라 2013년 200만호 AMI구축 1차 사업을 시작으로 2020년까지 2,250만호 고객 전체에 AMI를 구축한다는 목표로 2018년까지 약 680만호를 구축 완료했으며 2019년에도 400만호 설치를 계속 진행 중에 있다.

표 1. AMI 보급 전망(기존) 및 실적 (단위 : 만호, 누적)

연 도	~'15 (1차)	'16 (2차)	'17 (3차)	'18 (4차)	'19 (5차)	'20 (6차)
전 망	730	1,000	1,250	1,500	1,830	2,250
실 적	250	435	520	680	-	-

※산업통상자원부 보도자료(2018.7.18.)

AMI 보급이 100%로 완료되면 여러 가지 새로운 서비스들이 생겨나서 실생활에 도움이 되고 많은 변화가 일어날 것이다. 예로 전기 사용량 패턴분석을 통한 상점의 영업시간 예측서비스, 독거노인을 위한 생활안전서비스, TOU(Time Of Use) 적용으로 계시별 요금제 등 다양한 서비스들이 출현 할 것으로 예상된다.

* 정회원, 한전KDN 차장
** 준회원, 조선대학교 컴퓨터공학과 박사과정
*** 정회원, 조선대학교 컴퓨터공학과 교수
본 연구는 한국전력공사의 2018년 기초연구개발 과제 연구비에 의해 지원되었음(과제번호: R18XA06-16)

접수일자 : 2019년 08월 30일
게재확정일 : 2019년 12월 17일
교신저자 : 김판구, e-mail : pkkim@chosun.ac.kr

AMI데이터는 다양하지는 않지만 방대한 양이며 실시간적인 데이터라고 할 수 있다[1,2]. AMI를 이용한 서비스들을 제공하기 위해선 필수적으로 전력량계로부터 방대한 계량데이터들을 잘 취득하여야 한다. 그러나 국내AMI는 대부분이 PLC(Power Line Communication)방식을 사용하여 계량데이터를 취득하고 있다. PLC통신방식의 특징이 노이즈의 영향이 많아 통신이 잘 되지 않는다. 그래서 계량데이터 취득에 많은 애로사항이 존재한다. 현재 국내AMI 계량데이터 취득 성공률은 약 93~95% 수준에 불과하다.

본 논문에서는 누락된 계량데이터의 추정을 통해서 여러 가지 서비스에 필요한 기본데이터를 충실히 보충하고자 한다. 계량데이터 추정을 위한 선행 기술로는 빠진 데이터의 중간값을 계산하여 입력하는 방식과 사용량 패턴분석을 통해 사용량을 추정하는 방식이 있다. 그러나 첫 번째 방식인 중간값을 계산하는 방식은 중간 1건의 데이터 누락시 효과적이다. 다건이 누락됐을 경우에는 사용할 수 없다. 사용량 패턴 즉 평소 같은 시간대의 평균사용량 분석을 통한 추정방식은 평소 패턴과 다르게 국경일이나, 임시휴일, 날씨 등 예측이 어려운 사회적 현상일 경우 상당한 오차가 발생할 수 있다. 그래서 본 논문에서는 이 방식을 보완 할 수 있도록 유사도를 활용하여 사용량 패턴이 유사한 고객을 찾아 누락데이터를 보정하는 방식을 제안하고 선행 두 가지 방식과의 성능 비교 및 정확도를 계산하여 검증하고자 한다.

Ⅱ. 본 론

1. 관련 연구

가. 선행 기술 탐구

첫 번째 누락데이터 보정방법으로 중간값을 계산하여 추정하는 방법이다. 가장 일반적인 방법이다. 1건의 누락일 경우 효과적일수 있으나 다건의 누락일 경우 정확도가 떨어져서 사용 할 수가 없다.

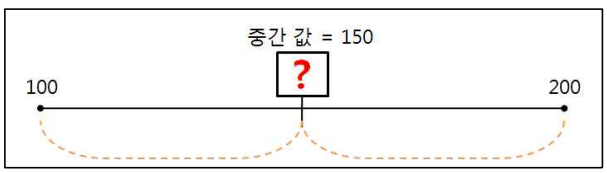

그림 2-1. 한건 누락데이터 보정

두 번째 누락데이터 보정방법으로 자신의 과거 전기사용량 패턴분석 즉 과거 동시간대 전기사용량 평균을 통한 사용량을 예측하여 추정하는 방법이 있다.

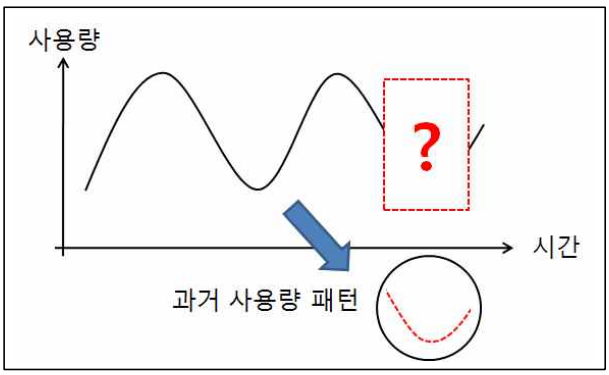

그림 2-2. 사용량 패턴 분석

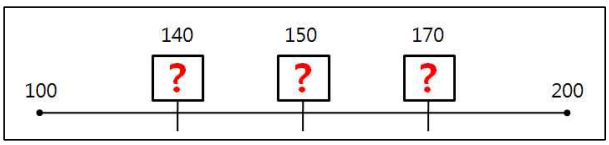

그림 2-3. 다건 누락 데이터 보정

그러나 이 방법은 국경일이나, 임시휴일, 날씨 등 과거 사용량 패턴과 다르게 불규칙적인 경우 상당한 오차가 발생 할 수 있다.

나. 유사도 계산

유사도 측정 방법에는 유클리디안, 맨하튼 거리, 피어슨 상관계수, 코사인 거리, 타니모토(자카드)거리 측정법 등이 있으며 가장 보편적으로 유클리디안 거리와 코사인 거리, 피어슨 상관계수, 타니모토(자카드) 거리 측정법이 활용되고 있다.

유클리디안 거리는 소비자 또는 아이템 간의 유사도를 계산하는 가장 쉬운 방법으로 가장 직관적이고 직관적인 거리의 개념이다. 일반적으로 2차원 상에서 두 점간의 거리는 피타고라스 정리에 따라 계산된다. 하지만 n차원의 공간에서 두 점간의 거리를 구하기 위해 피타고라스의 정리를 조금 더 확장시킨 것이 유클리디안 거리로 해당 수식은 그림 2-4와 같다.

$$EuclideanD(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

그림 2-4. 유리클리안 거리

여기서 (x, y)는 두 개의 연속적인 데이터이며, n은 데이터 세트의 자료의 개수를 의미한다. 그림 2-4를 통해 구해진 유클리디안 거리는 거리의 최댓값이 존재하지 않아 해당 거리를 비교할 수가 없다. 따라서 해당 거리 법을 쓰기 위해선 0과 1사이의 값으로 데이터의 정규화가 우선적으로 이루어져야 하며, 정규화 된 거리에서 두 벡터(Vector)가 가까울수록 0에 가깝고

멀수록 1에 가까워지게 된다[3]. 하지만 유클리디안 거리를 통한 유사도 계산 방식은 두 벡터간의 단순한 거리를 계산한다는 점에서 해당 벡터가 같은 방향성을 지니고 있는지를 확인할 수가 없다. 따라서 해당 유사도를 활용할 경우 두 벡터 간의 유클리디안 거리가 같다면 다른 방향성을 갖더라도 유사한 정도가 큰 것으로 나타날 수 있다는 한계점이 존재한다.

코사인 유사도는 내적공간에서 두 벡터 간의 각도를 코사인(Cosine)방식을 이용하여 측정한 값이다. 두 벡터 간의 각도가 0°로 그 방향이 완전하게 같다면 코사인 값은 1, 90°의 각도로 서로 관계가 없다면 0, 180°로 두 벡터간의 방향이 완전히 반대일 때는 -1과 같이 -1과 1의 사이 값을 갖게 되는데, 이때 코사인 유사도의 결과 값은 0과 1 사이의 양수 공간에서 표현되며 이 값은 벡터의 크기가 아닌 두 벡터간의 유사한 정도를 나타낸다. 코사인 거리의 계산식 그림 2-5는 다음과 같다[4].

$$x \cdot y = \| x \| \| y \| \cos\theta$$

$$CosSim(x,y) = \frac{x \cdot y}{\| x \| \| y \|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}}$$

그림 2-5. 코사인유사도 거리

여러 가지 유사도 측정 방법 중 코사인 유사도는 각도 값을 이용하여 유사한 방향으로 뻗어나가는지를 찾기 때문에 모든 벡터가 양수만을 가지고 있다고 간주하며 0에서 1사이의 값만 추출되고, 이는 정규화가 되어 있는 것으로 볼 수 있어 데이터를 별도로 정규화 시킬 필요가 없다. 또한 벡터 간의 양적 값을 이용해 거리를 계산하는 유클리디안 거리보다 비슷한 성향의 것을 찾아낼 수 있다는 점에서 유사도 측정에 많이 활용되고 있다. 하지만 A, B, C 벡터간의 방향성이 서로 같아 코사인 거리가 0일때, A가 B와 C중에 어느 벡터와 근접하고 있는지를 판별할 수는 없다는 한계점이 있다.

피어슨 상관 계수를 통한 유사도는 유클리디안, 코사인 거리 측정방법과 다르게 두 변수 간의 상관관계를 통해 얻어진다. 따라서 사용자 기반 협업 시스템에서 유사도 계산이 되는 대상은 사용자가 되며, 아이템 기반 협업 시스템에서는 두 개의 아이템이 계산 대상이 된다. 피어슨 상관계수는 두 변수(벡터)간의 공분산 값을 변수들의 표준편차의 곱으로 나눈 값으로 그 공식은 그림 2-6과 같다[5].

$$PearsonD(r) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

그림 2-6. 피어슨상관계수 거리[6]

피어슨 상관계수를 통한 유사도의 경우 두 변수간의 상관관계, 즉 두 사용자가 따로 변하는 정도를 두 사용자가 함께 변하는 정도로 나눈 값으로 나타나기 때문에 사용자가 어떠한 아이템에 대해 평점을 높게 측정하여도 해당 점수의 영향을 덜 받으며 해당 값이 1이라면 두 변수가 완전히 동일하다고 보며, 전혀 다르다면 0, 반대방향으로 완전히 동일하면 -1의 값을 갖는다[7].

3. 유사도를 활용한 누락데이터 보정

가. 알고리즘

누락데이터 보정방법으로 먼저 데이터 전처리를 통해 각 고객별 시간대별 사용량을 계산한다.

두 번째로 누락된 고객과 유사한 고객을 찾아서 유사고객 목록쌍을 생성한다. 이때 유사한 고객을 찾는 알고리즘으로는 유클리디안 유사도, 코사인 유사도, 피어슨 상관계수 등 3가지를 이용한다.

세 번째로 각 유사도에 의해서 만들어진 유사고객 목록쌍에 해당하는 고객의 전기사용량을 얻어온다. 그리고 누락된 시간의 데이터를 유사고객쌍의 동시간대 사용량 데이터를 이용하여 전시간대 누락고객의 누적사용량에 합산하여 보정한다.

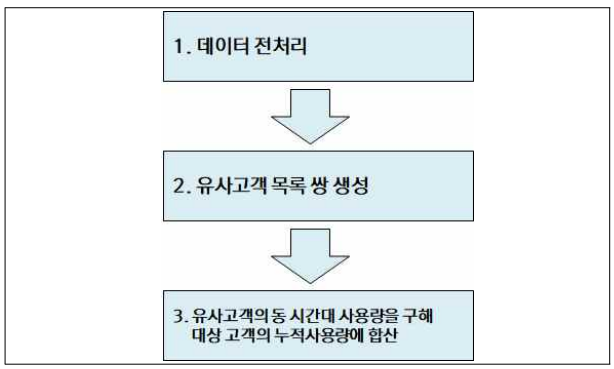

그림 3-1. 처리 알고리즘

나. 데이터 전처리

본 논문에서는 전기사용량 LP(Load Profile) 샘플 데이터를 준비하였다. 전력량계에서 수집된 데이터는 시간대별 사용량 데이터가 아니고, 계속 증가하는 누적사용량 데이터이다. 그래서 후 시간대 누적사용량에서 전 시간대 누적사용량 차이를 구해야 해당 시간대 사용량이 된다. 전처리 과정을 통해서 모든 고객의 사용량을 미리 계산해서 사용해야 한다. 모든 고객의 누적사용량은 차이가 많이 있다. 고객별 유사도를 판별하기 위해서 누적사용량이 필요한 것이 아니고 시간대별 사용량이 필요하고 이 데이터를 사용하여 유사고객 선정한다.

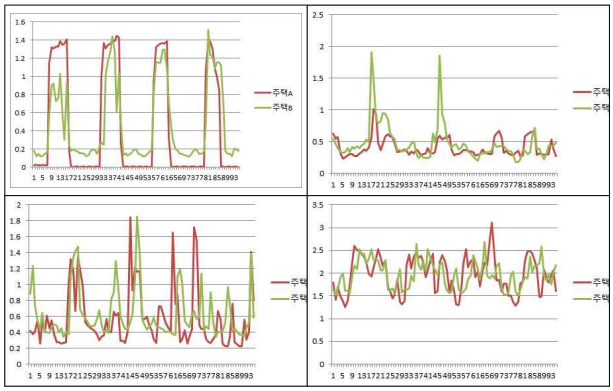

그림 3-2. 유사도계산결과 데이터 비교

다. 실험 및 평가

본 논문에서는 약 1천만건의 가상의 전기사용량 데이터 샘플을 사용했다. 이중에서 100건의 고객을 랜덤하게 선정하여 이 고객들이 누락데이터가 발생했다고 가정하고 두 가지 방법으로 실험을 진행했다.

첫 번째는 1건의 데이터가 누락됐을 때 비교실험이다. 먼저 과거 방식은 전.후 데이터 차이의 중간 값을 계산하여 보정하는 방식과 유사도를 계산하여 유사고객을 선정하고 유사고객의 동시간대의 사용량을 이용하여 보정하는 방식이다. 이때 유사도 계산에 사용하는 알고리즘은 3가지를 사용했다. 아래 그림은 1건의 데이터 누락 보정에 대한 3가지 방식의 비교 실험한 결과이다.

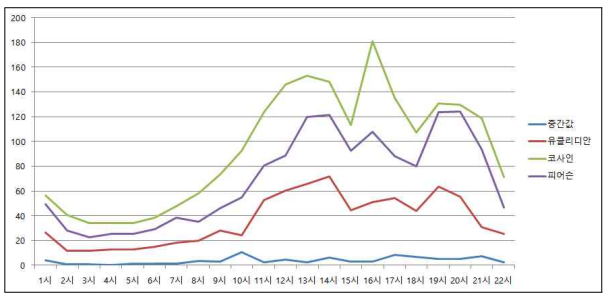

그림 3-3. 1건 누락 보정 SSE(Sum of Squares Error)

위 그림에서 보듯이 한건 누락에 따른 누락데이터 보정했을 때 중간값을 보정하는 것이 실데이터와 비교시 SSE값이 가장 낮아 성능이 좋음을 알 수 있다.

두 번째는 다건 데이터가 누락됐을 때 비교실험이다. 과거 방식은 전일분 동시간대 사용량을 계산하여 누락시간대의 데이터를 추정하는 방식이고, 본 논문에서 제시안 유사도를 계산 알고리즘 3가지를 이용하여 유사고객을 선정하고 유사고객의 동시간대의 사용량을 이용하여 보정하는 방식이다.

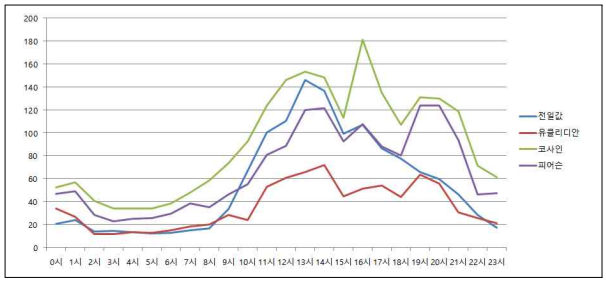

그림 3-4. 다건 누락 보정 SSE(Sum of Squares Error)

위 그림에서 보듯이 유클리디안 유사도를 이용한 값의 오차합이 코사인 유사도와 피어슨 상관계수 알고리즘을 이용한 값보다 월등히 좋음을 알 수 있다.

III. 결 론

각 가정의 전력량계 데이터를 취득하는 AMI시스템은 대부분이 PLC통신방식을 사용하기 때문에 회선품질이 좋지 못하다. 그래서 데이터 취득의 누락이 많이 발생하고 있다. 본 논문에서는 AMI시스템의 효용성을 올리고자 유사도를 활용한 누락데이터 보정하는 방법을 제안했다.

한 건의 누락데이터 발생에 따른 보정방법은 유사도를 활용한 방법보다는 전.후 데이터의 중간값으로 보정하는 것이 실험을 통해 증명되었다.

연속된 다건의 누락데이터 발생에 따른 보정방법은 여러 가지 실험을 통해 과거 사용량 패턴과 유클리디안, 코사인, 피어슨 상관계수 등 유사도 알고리즘를 활용한 방법으로 비교시험한 결과 유클리디안 유사도를 활용하여 보정하는 방법이 좋은 성능을 보였다.

결론적으로 유사도를 활용한 방법이 날씨, 기온, 요일, 휴일 등 불규칙적인 전력사용패턴을 보일 경우 더 유리할 것으로 보이며, 앞으로 유사도 판별방법으로 유클리디안 방법외에 다른 여러 가지 방법을 추가 연구하여 더 유사한 고객을 찾아 그 고객을 사용량을 누락데이터 보정에 활용하고자 한다.

REFERENCES

[1] Choong Kwon Lee, "A Study of Big Data InformationSystems Building and Cases," *Smart Media Journal*, vol. 4, no. 3, pp. 56-61, 2015.

[2] Tae Woong Kim, "Group Behavior Pattern and Activity Analysis System Using Big Data Based Acceleration Signals," *Smart Media Journal*, vol. 6, no. 3, pp. 83-88, 2017.

[3] 데이터 분석에서 나오는 수학 - 유클리디안 거리. http://egloos.zum.com/metashower/v/9957577 (accessed Aug., 24, 2019).

[4] 유사도-위키백과. https://ko.wikipedia.org/wiki/%EC%BD%94%EC%82%AC%EC%9D%B8_%EC%9C%A0%EC%82%AC%EB%8F%84 (accessed Aug., 24, 2019).

[5] 이성현, "R 패키지 Recommenderlab을 이용한 추천 시스템 성능평가", *동국대학교 학사 졸업논문*, 2015. 2

[6] Byung-Ik Ahn, Ku-Imm Jung, Hae-Lim Choi, "A Study on Recommendation Systems based on User multi-attribute attitude models and Collaborative filtering Algorithm," *Smart Media Journal*, vol. 5, no. 2, pp. 84-89, 2016.

[7] 상관분석 유사도 - 3. 피어슨 유사도. http://bigBigdata.tistory.com/99?category=529087 (accessed Aug., 24, 2019).

저 자 소 개

권혁록(정회원)

1997년 경일대학교 컴퓨터공학과 학사 졸업(공학사)
2018년 조선대학교 소프트웨어융합과 석사 졸업(공학석사)
2018년~현재 조선대학교 컴퓨터공학과 박사과정
1996년~현재 한전KDN 재직중

<주관심분야 : AMI, 딥러닝>

홍택은(준회원)

2015년 조선대학교 컴퓨터공학부 졸업(공학사)
2017년 조선대학교 소프트웨어융합과 석사 졸업(공학석사)
2017년~현재 조선대학교 컴퓨터공학과 박사과정

<주관심분야 : 지능형정보처리, 자연어처리, 딥러닝>

김판구(정회원)

1988년 조선대학교 컴퓨터공학과 졸업(공학사)
1990년 서울대학교 컴퓨터공학과 석사 졸업(공학석사)
1994년 서울대학교 컴퓨터공학과 박사 졸업(공학박사)
2007년~현재 조선대학교 컴퓨터공학과 교수

<주관심분야 : 지능형정보처리, 시맨틱 웹, 온톨로지, 자연어처리, 데이터 마이닝 등>