J Korea Industr Inf Syst Res Volume 18 Number 3
http://dx.doi.org/10.9723/jksiis.2013.18.3.035

센서스 정보 및 전력 부하를 활용한 전력 수요 예측†

(Forecasting Electric Power Demand Using Census Information and Electric Power Load)

이 헌 규*, 신 용 호**
(Heon Gyu Lee and Yong Ho Shin)

요 약 국내 전력 수요량 예측을 위한 정확한 분석 모델을 개발하기 위하여 고차원 데이터 군집 분석에 적합한 차원 축소 개념의 부분공간 군집 기법과 SMO 분류 기법을 결합한 전력 수요 패턴 예측 방법을 제안하였다. 전력 수요 패턴 예측은 무선부하감시 데이터 뿐 아니라 소지역 단위의 센서스 정보를 통합하여 시간대별 전력 부하 패턴 분석과 인구통계학 및 지리학적 특성 분석이 가능하다. 서울지역 대상의 센서스 정보 및 전력 부하를 이용한 소지역 전력 수요 패턴 예측 결과 총 18개의 특성 군집을 구성하였으며, 전력 수요 패턴 예측 정확도는 약 85%를 보였다.

핵심주제어 : 부분공간 군집화, 센서스 정보, 전력 수요 예측, 소지역 군집, 데이터 마이닝

Abstract In order to develop an accurate analytical model for domestic electricity demand forecasting, we propose a prediction method of the electric power demand pattern by combining SMO classification techniques and a dimension reduction conceptualized subspace clustering techniques suitable for high-dimensional data cluster analysis. In terms of electricity demand pattern prediction, hourly electricity load patterns and the demographic and geographic characteristics can be analyzed by integrating the wireless load monitoring data as well as sub-regional unit of census information. There are composed of a total of 18 characteristics clusters in the prediction result for the sub-regional demand pattern by using census information and power load of Seoul metropolitan area. The power demand pattern prediction accuracy was approximately 85%.

Key Words : subspace clustering, census information, power demand forecasting, microarea clustering, data mining

1. 서 론

전력 수요량 예측을 위한 정확한 분석 모델은 전력 운영과 계획에 필수적이며, 전력 산업에 있어서 전력의 구입, 생산, 로드 스위칭, 그리고 기반 시설의 업데이트 등에 대한 의사 결정을 함에 있어서 도움을 준다. 특히, 전력 수요 예측은 공급자, 독립계통운영회사 및 재무기관이 전기에너지 생산, 전송, 분배, 및 관련 마케팅에 있어서 매우 중요하다. 전력 수요는 계절과 시간의 변화에 따라 크게 변동하는 양태를 보이며, 이

† 이 논문은 지식경제부 우정사업본부의 우정기술연구개발사업 (2006-X-001-02, SMART Post 구축 기술 개발) 지원으로 수행되었음
* 한국전자통신연구원 융합기술연구부문, 제1저자
** 영남대학교 경영학부, 교신저자(yhshin@yun.ac.kr)

로 인해 전력수요가 일시적으로 크게 변할 수 있으므로 저장 불가능한 전력은 공급과 수요가 항상 일치하여야 한다. 정확한 전력 수요의 예측하는 것이 합리적인 전력 공급 계획을 수립 하는 데에 필수불가결한 조건이다.

국내 전력 수급 위기 상황으로 2011년 9.15 전력순환 단전 사태가 발생하였고 2012년에는 예비력 400만 kW 미만의 수급비상단계인 '관심' 단계 이하가 10회, 예비력 500만 kW 이하인 '준비' 단계 발령까지 포함하면 총 52 일간 수급비상 상황이 발생했다 [1]. 최근 이러한 전력 수급비상 상황에 대해서 전력수급 안정을 위해서는 현재 고압 고객인 산업체에 집중된 수요관리를 확대하여 상가, 사무용빌딩, 아파트 등의 저압 고객까지 수요량 관리를 확대할 필요가 있으며, 보다 정확한 수요 예측을 위해서 예측시스템 업그레이드 및 부하(load) 패턴 분석을 강화할 필요가 있다.

전력 산업에 있어서 효율적인 운용과 수급 계획을 위한 수요 패턴 예측 기술로는 통계 및 데이터마이닝 등과 같은 수학적 방법들이 수요량 분석 모델링을 위해 사용된다. 수요 패턴 예측 과정은 기존의 수요 패턴을 식별하고 통계적 분석 [2] 및 데이터마이닝 [3], [4] 기술을 적용하여 새로운 수요량을 예측 하는 것을 말한다. 일반적으로, 데이터마이닝 기술을 적용한 전력 수요패턴 예측은 관련된 정보로부터 전력 패턴 모델을 생성하고 이 모델을 적용하여 새로운 부하패턴을 예측한다. 그러나 기존의 연구에서는 저압 고객과 같은 다양한 계약종별 대상의 수요 예측이 아닌 전력 수요량이 큰 산업체의 고압 고객만을 대상으로 하고 있다. 저압 고객의 경우는 무선부하감시 장치 또는 변압기를 통해 수집된 전력부하이다. 이는 예측 모델을 생성하기에 너무 작은 단위의 상세 측정 부하(1가구 단위 또는 1개의 변압기 설치 단위)이므로 수요량 예측의 정확성을 보장할 수 없다. 또한 행정구역 단위로 측정값을 집계할 경우, 범위가 넓고 너무 일반화(generalization)된 정보이므로 예측 모델 적용이 불가능하다. 따라서 저압 고객의 다양한 계약종별에 대한 전력 수요량 예측 모델의 underfitting 및 overfitting을 방지 하면서 정확성을 높일 수 있는 새로운 측정 단위의 집계가 필요하다.

이 논문에서는 전력 수요량 예측시스템에 적용 가능한 저압 고객 대상의 군집화 기반 예측 기법을 제안한다. 이를 위해서 변압기 전력 부하 데이터를 행정구역 단위보다 작은 통계청 인구조사 단위인 소지역(집계구) 공간 개념을 적용하며, 예측 모델 생성을 위한 알고리즘으로 다차원 속성의 동적 군집화 및 분류가 동시에 적용되는 부분공간 군집화 기법을 확장하여 적용한다.

<그림 1>은 논문에서 제안하는 군집/예측 기법의 프로세스이며 상세 내용은 다음과 같다.

<그림 1> 전력 수요 예측 프로세스

- 데이터 전처리 및 통합
  - 통계청 자료인 인구/가구/주택/사업체 관련 '센서스 공간 DB'와 한전전력연구원 저압고객 변압기 부하 분석 GIS 시스템으로부터 센서스 정보 및 변압기 시간별 전력 부하 데이터를 수집
  - 공간 연산 및 부하량 집계를 통해 소지역 단위의 인구통계학적 속성 데이터와 전력 부하 패턴을 추출

- 부분공간 군집화
  - 고차원 데이터의 희소성 문제 해결과 효과적인 군집화를 위해 관련성 높은 부분차원을 찾아내어 군집을 구성하는 cell-based, density-based, cluster-oriented 부분공간 군집화 알고리즘의 적용

- SVM (Support Vector Machine) 기반 예측 모델
  - SMO (Sequential Minimal Optimization) 분류 기법을 통한 소지역 단위 전력 수요 패턴 예측

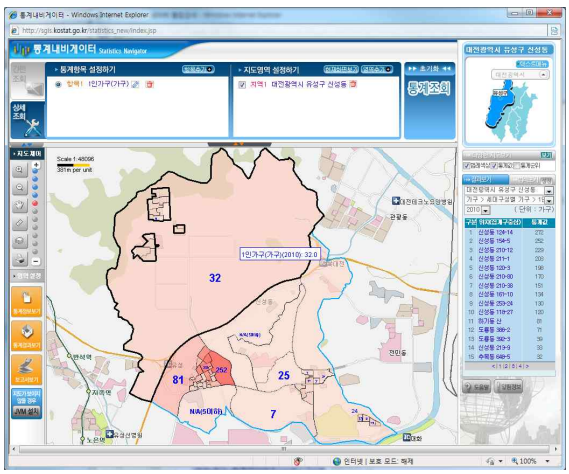
및 센서스/시간 특성 분석
- 정확도 평가 항목을 통한 최적 부분공간 군집화 기법 선택 및 소지역 군집화

2. 관련 연구

2.1 통계청 센서스 공간 DB

통계지리정보 서비스의 센서스 및 공간 데이터는 국내 인구, 가구, 주택 및 사업체 총조사의 '센서스 공간 DB [5]'[1]을 구축하여 다양한 수요기관에 제공하는 지리정보 기반의 통계 서비스이다. 자료 제공 범위는 소지역 집계구[2] 단위로 제공하며, 통계적 노출 관리 기법을 적용한 마이크로 데이터를 제공하므로 소지역 단위의 인구 통계학적 데이터 이용이 가능하다. GIS 기반의 의사결정 도구로서 '통계 내비게이터' 서비스를 제공하여, 약 1,350만개의 거처와 320만개 사업장에 대한 위치정보와 통계를 결합하여 구축한 DB를 기반으로 시도/군구/행정동/소지역 단위의 영역 설정을 통한 통계 정보를 제공한다.

이 정보는 마케팅 목적의 고객 특성과 지역적 특성에 맞는 마케팅 전략 수립의 과학적인 의사결정 지원 도구로 활용 가능하다. (<표 1>과 <그림 2> 참조)



<그림 2> 통계청 통계내비게이터 실행 예

<표 1> 센서스 자료 제공 내용 및 범위

통계 항목		세부 항목
대분류	소분류	
총괄	인구총괄	총인구, 평균나이, 인구밀도, 노령화지수, 노년부양비, 유년부양비, 총부양비
	가구총괄	총가구, 가구별 점유형태(자가), 가구별 점유형태(전세), 가구별 점유형태(기타), 평균가구원
	주택총괄	총주택
	시설물총괄	아파트(거처), 단독주택 거처, 기타거처, 총괄사업체수
인구	성/연령별 인구	4세 이하, 5세~9세, 10세~14세, 15세~19세, 20세~24세, 25세~29세, ..., 65세 이상
	교육정도별 인구	초등학교, 중학교, 고등학교, 대학교, 대학원
	성/혼인 상태별인구	미혼_남자, 배우자 있음_남자, 사별_남자, 이혼_남자, 미혼_여자, 배우자 있음_여자, 사별_여자, 이혼_여자
	종교별인구	종교 있음, 불교, 기독교(개신교), 기독교(천주교), 유교, 원불교, 증산교, 천도교, 대종교, 기타
가구	세대구성별 가구	1세대, 2세대, 3세대, 4세대, 1인 가구, 비혈연가구
	점유형태별 가구	자가, 전세, 월세, 사글세, 무상
	난방시설별 가구	중앙난방, 지역난방, 도시가스보일러, 기름보일러, 프로판가스보일러, 전기보일러, 연탄보일러, 연탄아궁이, 재래식아궁이, 기타
	가구 실태	방1개, 방2개, 방3개, 방4개, 방5개, 거실없음, 거실1개, 거실2개, 식당없음, 식당1개, 식당2개
주택	건축연도	건축연도별 주택수
	건평	7평 미만, 7~9평, 10~14평, 15~19평, 20~29평, 30~39평, 40~49평, 50~69평, 70평 이상
	주택유형	다세대, 단독주택, 아파트, 연립주택, 영업용 건물 내 주택, 주택이외 거처
사업체		사업체수, 종사자수

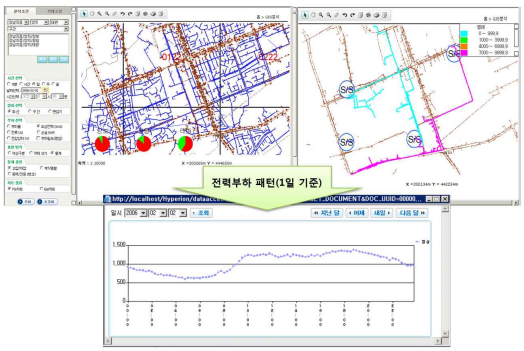
현재 인구/가구/주택은 2010년도 기준이며, 사업체는 2011년 기준인 센서스 조사 정보이다. 전력 수요량

1) 인구주택 및 사업체 센서스 자료와 이에 대응하는 위치정보를 부가한 센서스 개별 공간 DB와 센서스 지도정보, 센서스 경제정보가 결합된 통계 DB
2) 통계청에서 60가구 약 인구 500명 단위의 센서스 조사구

예측의 인구통계학적, 지리학적 데이터 추출을 위해서는 통계청 통계지리정보시스템의 오픈 API 서비스 신청으로 센서스 속성 데이터 및 공간데이터(전자지도) 획득이 가능하다.

2.2 GIS 기반 전력부하 분석 시스템

한전 전력연구원에서는 인터넷 환경에서의 배전 계통 GIS 구축을 통한 전력부하 분석 시스템을 개발하였다. 이 시스템은 원격검침 수요정보 및 전력 부가서비스 모델 개발을 목적으로 고객/검침/부하관리/분석 기능, 부하계산 결과데이터 검증, 공간적 분포 특성을 고려한 검침데이터 현황 및 통계서비스 모듈 및 배전계통 운영 효율화를 위한 부하변경 시뮬레이션 등을 포함한다.



<그림 3> GIS 기반 전력부하 분석 시스템

<그림 3>의 전력 부하 분석 시스템은 부하패턴 통계 분석용 OLAP 및 배전계통 GIS에서 회선/구간/변압기 위치 확인 및 부하 현황/밀도 파악이 가능하며, 회선 또는 특정 구간의 당일/월간(시점/최대부하)/연간 전력 부하 통계 분석을 지원하는 시스템이다. 그러나 이 시스템은 과거부터 현재까지의 원격검침 장비를 통해 수집된 전력 부하량을 GIS와 연동하여 통계분석을 지원하므로 향후, 전력 수요량 변화 및 부하 패턴 분석/예측의 예측 시스템 기능을 포함하지 않는다.

이 논문에서는 <그림 3>의 시스템으로부터 저압 고객 변압기 전력 부하 데이터와 GIS를 통한 변압기 위치 정보를 수집하여 전력 수요 예측 기법에 적용한다.

2.3 부분공간 군집화 기법

부분공간 군집화는 데이터의 서로 다른 부분속성들을 군집화에 유용한 특징벡터만을 동적으로 구성하여 그룹화하는 기법이다. 즉, 모든 데이터 차원의 부분집합에서 유사도가 가장 높은 군집을 찾는 기법이다. 부분공간 군집화는 군집 정의 및 군집 생성 방법에 따라 3개의 주요한 기법으로 구분된다 [6]. 첫 번째 기법은 cell-based 방식으로 데이터 공간을 격자 셀로 나누고 충분한 밀도를 가지는 셀들로부터 군집을 형성하는 방식이다. 기본적인 개념은 먼저 격자 셀 집합을 정의한 후에 객체들을 적절한 셀에 할당하고 각 셀의 밀도를 계산한다. 다음으로 특정 임계치 이하의 밀도를 갖는 셀을 제거한고 밀도가 높은 연속한 셀들로부터 군집을 형성한다. 대표적인 cell-based 방식으로 CLIQUE [7]는 부분공간의 군집들을 절차적으로 찾는 격자 기반의 군집화 알고리즘이다. 이 방법은 밀도기반 군집화의 단조성을 활용하여 k차원(속성)에서 밀도기반의 군집을 형성하는 점들은 이차원의 모든 가능한 부분집합에서도 특정 밀도 기반 군집의 부분을 형성한다는 성질을 이용한다.

두 번째 부분공간 군집화 기법은 density-based 방식이다. 이 방법은 낮은 밀도의 지역에 의해 분리된 높은 밀도의 지역들을 파악함으로서 군집을 형성시킨다. 이 논문에서 사용된 density-based 알고리즘은 필터-정제의 2단계 구조를 갖는 FIRES [8] 이다.

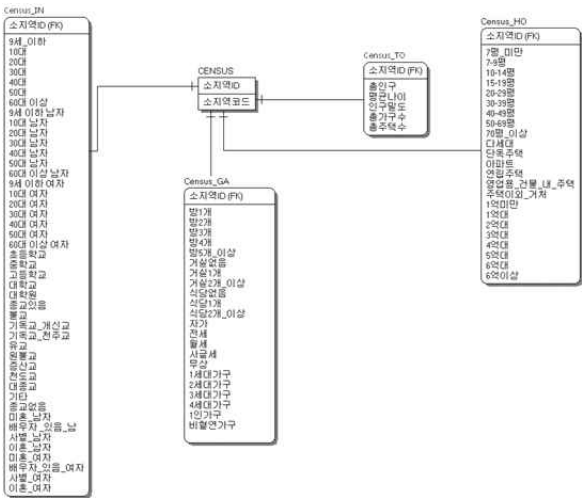
세 번째, 군집 방법은 clustering-oriented 기법이다. 고차원 데이터에 대한 군집화에서 데이터의 차원이 증가할수록 데이터의 희소성 문제 때문에 모든 차원을 고려하여 군집을 구성하면 성능이 현저히 떨어진다. 대표적인 알고리즘인 PROCLUS [9]는 차원-감소 부분공간 군집화 방법이다. 단일 차원 공간에서 출발하는 대신에 이 알고리즘은 고차원 공간에서 군집에 대한 초기 추정을 탐색하는 것으로 시작한다. 각 군집에 대해서 각 차원별로 가중치가 부여되고 이와 같이 갱신된 가중치는 다음 반속시에 군집을 재생성 하는데 사용된다.

이 논문에서는 3가지의 주요 기법의 대표 알고리즘인 CLIQUE, FIRES, PROCLUS를 전력 수요량 군집분석을 위해 적용하며, 실험을 통해 비교 평가한다.

3. 데이터 전처리 및 통합

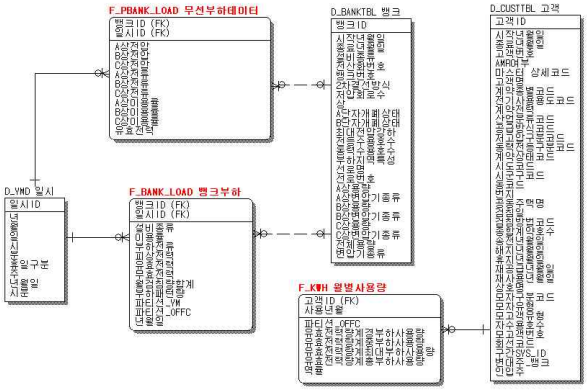
한전의 GIS 전력부하 데이터 수집 및 분석 시스템은 현재 서울지역을 대상으로 구축되어 있으므로 전력 부하 및 센서스 데이터 수집 범위를 서울지역으로 한정한다.

통계청으로부터 획득한 <표 1>의 센서스 정보 중 사업체를 제외한 인구/가구/주택/총괄 부문 속성 정보와 소지역 경계 및 행정구역 경계 전자지도를 사용한다. 센서스 정보에 대한 ER-다이어그램은 <그림 4>와 같다.



<그림 4> 센서스 정보에 대한 ER-다이어그램

한전의 무선부하감시 시스템에서 측정한 변압기(Bank) 정보는 30분 간격으로 전력 부하가 그림 5의 테이블 형태로 저장된다. 이 중 전력 수요량 예측을 위해서 필요한 정보들을 추출한다. 먼저 변압기 코드 집합을 생성을 위한 D_BANKTBL 테이블의 'Bank ID'와 '총 사용량' 고객 정보(D_CUTTBL) 테이블의 '계약전력' 속성 정보를 추출하고 월별사용량(F_KWH) 테이블에서 '유효전력량계총부하사용량'을 추출한다. 변압기 정보 테이블과 고객 정보 테이블의 조인을 위해서는 변압기(Bank) 정보 테이블에서 '시작년월일', '종료년월일', '전산화번호' 속성과 고객 정보 테이블에서 '변대주_뱅크' 속성을 이용한다. 변압기별 30분 단위의 전력 부하 패턴을 생성하기 위해서는 무선부하 데이터(F_PBANK_LOAD) 테이블에서 '유효전력'과 '일시 ID' 속성을 추출한다. 또한, 무선부하감시 시스템에 의해 측정된 변압기 유효 전력의 시간 정보를 추출하기 위하여 일시(D_YMD) 테이블의 '년, 월, 일, 시, 분' 속성을 이용한다.



<그림 5> 변압기 부하 분석 대상 테이블 ER-다이어그램

최종 추출된 전력 부하 관련 데이터집합은 <표 2>와 같다.

<표 2> 전력 수요량 예측을 위한 변압기 정보

속성명	데이터 타입	속성 설명
BANK_ID	nominal	변압기 구분 코드
CAPA	nominal	총용량
CNTR_KND_CD	nominal	계약 종별
WHM_TOT_USE KWH	continuous	유효 전력량계 총부하사용량
KW	continuous	유효전력 (월별 30분 간격)

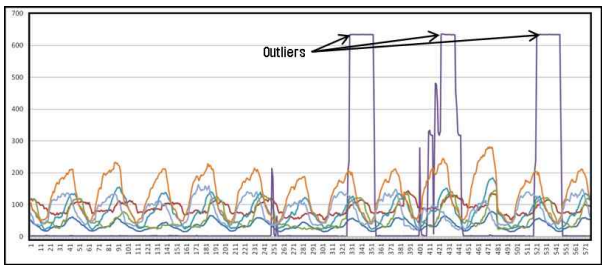
변압기 전력 부하 데이터로부터 전력 부하 패턴을 생성한다. 부하 패턴은 30분 간격의 유효전력의 시계열 데이터로 일/주/월/년별로 생성 가능하다. 만약 월별 부하 패턴 생성할 경우, 다음 (식 1)과 같이 계산한다.

$$V^{(m)} = \sum_{i=1}^{30} \{V_0^{(m)}, \ldots, V_h^{(m)}, \ldots, V_H^{(m)}\}_i \quad \text{(식 1)}$$

여기서, m=소지역 ID, h(30분 간격의 부하 측정 시간)=0,...,1440, i(day)=1,...,30이다.

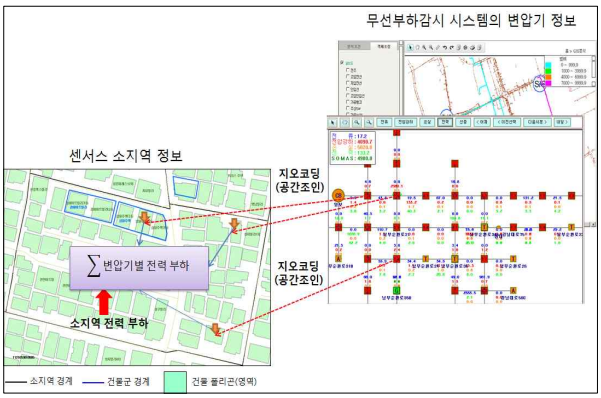
기존의 전력 수요 예측 연구에서는 고차원 데이터의 축소를 위해서 부하 패턴으로부터 특정 구간대의 정보만을 추출하여 새로운 특징벡터 [10]로 사용하였으나 이럴 경우, 군집화 전에 미리 특정 속성들만을 선택하기 때문에 정보의 손실을 가져올 수 있다. 따라서 이 연구에서는 군집화 알고리즘 자체에서 상관관계가 높은 특징들을 동적으로 선택하는 부분공간 탐색 방식을 적용하므로 30분 간격의 부하 패턴 모두를 특징벡터로 사용한다.

전력 부하 패턴은 무선부하감시 시스템의 변압기에서 측정된 정보이다. 이 데이터에는 측정 장치의 오류 또는 데이터 저장 단계에서의 오류가 포함되어 있다. 따라서 데이터 통합 전에 이러한 이상치(outlier) 데이터를 제거해야 한다. 이상치 데이터 발견은 전처리 단계의 군집화를 수행하여 군집 형성이 되지 않는 데이터를 삭제한다. 이상치 부하 패턴은 30분 간격의 모든 속성값을 입력하므로 부분공간 군집방식이 아닌 모든 차원을 고려하는 군집화 기법을 적용한다. 대표적인 군집화 알고리즘인 SOMs [11]를 적용하여 이상치 데이터를 제거를 위한 전처리를 수행한다. SOMs 군집화 전처리 기법은 코호넨 네트워크 모델을 적용하며, 구매 매트릭스는 12 by 12 인 군집수 144개로 설정하였다. <그림 6>은 전력 부하 패턴의 이상치 데이터의 예를 보여준다.



<그림 6> 부하 패턴에 포함된 이상치 데이터

센서스 정보와 변압기 전력 부하 패턴의 통합은 GIS 소프트웨어를 이용한다. 변압기 정보는 포인트 정보이며, 소지역은 폴리곤 정보이므로 공산연산자 중 공간 조인을 통해 수행한다. 하나의 폴리곤 형태의 소지역에는 다수의 변압기가 속하는 경우, 30분 간격의 부하 패턴 유효전력 값의 총합계를 구하여 소지역 단위의 전력 부하 패턴으로 계산한다. 또한 소지역에 변압기를 포함하지 않을 경우에는 그 소지역은 군집화 대상에서 제외시킨다. 센서스 정보와 변압기 전력 부하 패턴의 통합 과정은 <그림 7>과 같다.



<그림 7> 소지역 단위 데이터 통합 과정

4. 부분공간 군집화 기법을 이용한 전력 수요 예측

4.1 소지역 분류를 위한 부분공간 군집화

이상치 제거를 통한 전처리 및 통합된 데이터는 센서스 관련 96개 차원과 30분 간격의 월단위 변압기 부하 패턴의 유효전력인 1,440 차원을 고려해야 한다. 따라서 효율적인 소지역별 전력 수요 패턴 군집을 위해서는 고차원 부분공간 선택 군집 기법이 필요하며, 2장에서 설명한 cell-based, density-based, clustering-oriented 기법들의 대표 알고리즘을 적용한다. 부분공간 군집화 알고리즘은 Java Weka 프로젝트의 고차원 데이터를 위한 openSubspace: Weka Subspace-Clustering Integration [12]에서 오픈소스로 제공된다. openSubspace는 크게 3개 부분으로 구성된다. 부분공간 군집화 3개 기법인 cell-based, density-based, clustering-oriented 에 대해, 세부 10개 알고리즘을 제공하며, 각 알고리즘별 파라미터 셋팅을 포함한다. 또한 알고리즘 평가 기술로 F1-value, entropy/coverage, accuracy, CE/RNIA 방법을 제공한다.

이 논문에서는 유사한 전력 수요 패턴을 갖는 소지역들을 군집화 하고 새로운 전력 부하 패턴이 어느 그룹에 속하는지를 분류하는 예측 기법을 포함한다. 따라서 openSubspace에서 제공하는 평가 기술 중

accuracy 항목에 초점을 둔다. openSubspace의 accuracy 성능 평가는 군집화 수행 후에 각 군집 라벨을 분류 문제의 예측 대상인 클래스로 간주한다. 다음으로 결정 트리 기법을 이용하여 훈련 데이터를 입력 받아 해당 클래스를 정확히 예측하는지 평가한다. 그러나 결정트리 기법은 다량의 데이터 처리 문제와 센서스 통계 수치 및 부하 패턴과 같은 연속적(continuous) 수치값 처리에 성능이 좋지 못하므로, openSubspace에서 제공된 알고리즘의 평가 단계인 예측 기법 적용 부분에 대해 SVM [13], [14] 기반의 알고리즘으로 확장 적용한다. SVM 방식의 확장으로 단순히 부분공간 알고리즘들의 자체 성능 평가가 아닌 새로운 부하 패턴에 대한 정확한 군집 구성을 기대할 수 있다.

- CLIQUE 부분공간 군집화

CLIQUE는 2단계로 고차원 군집화를 수행한다. 첫 번째 단계에서, CLIQUE는 n차원 데이터 공간을 중첩하지 않는 직사각형의 단위 영역으로 분할하여 조밀 영역을 탐색한다. 이것을 각 차원(1-차원)들에 대해 수행한다. 이러한 조밀 단위 영역을 나타내는 부분공간들은 중첩하여 더 고차원의 조밀 단위영역들이 존재할 가능성이 있는 후보 탐색 공간을 형성한다. 두 번째 단계에서, CLIQUE는 각 군집에 대한 최소 포함을 생성한다. 각 군집에 대하여 연결된 조밀 단위영역들의 군집을 포함하는 최대 영역을 결정한다. 그 때 각 군집들에 대하여 최소 포함을 결정한다.

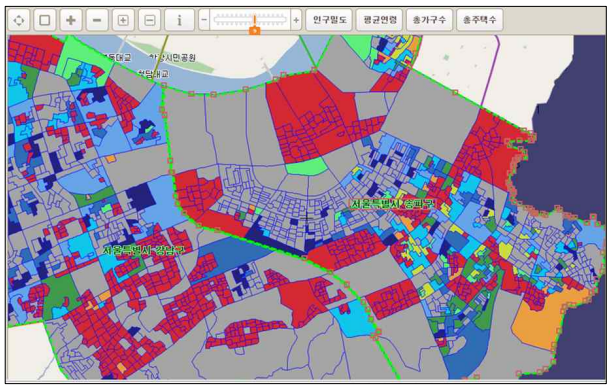
- FIRES 부분공간 군집화

FIRES는 filter-refinement 구조에 기반하며, pre-clustering, 부분공간 군집 근사들의 생성, post-processing 등 3단계를 거쳐 군집을 형성한다. 첫 단계에서 기본 군집(base cluster)이라 불리는 pre-clustering인 모든 1차원 군집을 계산한다. 다음으로, 부분공간 군집 근사 (approximation)들을 생성하며, 이것은 최대 고차원의 부분공간 군집 근사들을 찾기 위해, 기본 군집들을 병합시키는 과정이다. 마지막으로 부분공간 군집들의 후처리 단계에서는 군집 근사들을 정제하는 단계로, 의미 없는 기본 군집들을 식별하고 제거함으로서 후보 군집집합을 정제한다. 또한 데이터의 잡음을 제거함으로서 부분공간 군집을 완성한다.

- PROCLUS 부분공간 군집화

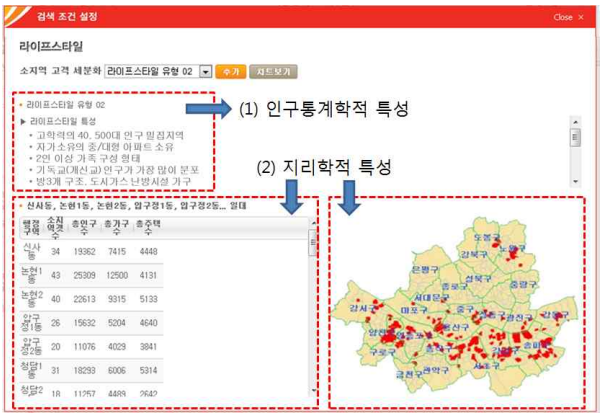
PROCLUS 알고리즘은 초기화, 반복, 정제의 3단계로 구성된다. 초기화 단계에서는 각 군집이 최소한 하나의 객체에 의해서 대표될 수 있도록 충분히 떨어진 초기 대표 객체인 medoid 집합을 선택하기 우해서 greedy 알고리즘을 사용한다. 생성하고자 하는 군집 수에 비례한 데이터 점들을 임의로 선택한 후에 greedy 알고리즘을 적용하여 다음 단계를 위해서 더 작은 최종 부분집합을 얻는다. 반복 단계에서는 감소된 집합으로부터 k개의 medoid를 임의로 선택한 후에 선정된 새로운 medoid 객체가 군집의 성능을 향상 시킬 경우, 이전의 medoid를 새로운 medoid로 교체한다. 각 medoid에 대해서 통계적 기대값보다 상대적으로 작은 평균 거리를 가지는 차원의 집합이 선택된다. medoid에 연관된 차원의 총 개수는 l이 군집 부분공간의 평균 차원 수를 선정하는 입력 모수인 경우 k*l 이다. 정제 단계에서는 탐색된 군집에 기반하여 각 medoid의 새로운 차원을 계산하여 점들을 medoid에 재할당하고 이상치들을 제거한다.

3가지 대표 부분공간 군집화 수행 결과 GIS를 이용하여 가시화가 가능하다. <그림 8>은 FIRES 알고리즘 적용 결과 서울의 해당 소지역을 총 18개로 군집화한 예를 보여준다. 또한 <그림 9>는 서울지역 특정 군집의 센서스 정보인 인구통계학적 특성 및 지리학적 특성(군집 내 소지역 분포) 특성을 나타낸다.



<그림 8> GIS 기반 소지역 군집 결과 (FIRES)

군집화 결과, 인구통계학 및 지리학적 특성뿐만 아니라, 각 군집의 대표적인 전력 수요 패턴을 추출할 수 있다. 이러한 대표 패턴은 각 군집의 모든 전력 부하 패턴의 평균으로 표현된다. <그림 10>은 3가지 부



<그림 9> 군집의 인구통계학/지리학적 특성

분공간 군집화 수행 후의 군집별 대표 전력 수요 패턴을 보여 준다.

따라서, 부분공간 군집화 결과는 소지역 그룹 단위 센서스 정보의 인구통계학/지리정보학적 특성 분석과 동시에 대표적인 전력 수요 패턴을 분석할 수 있다.

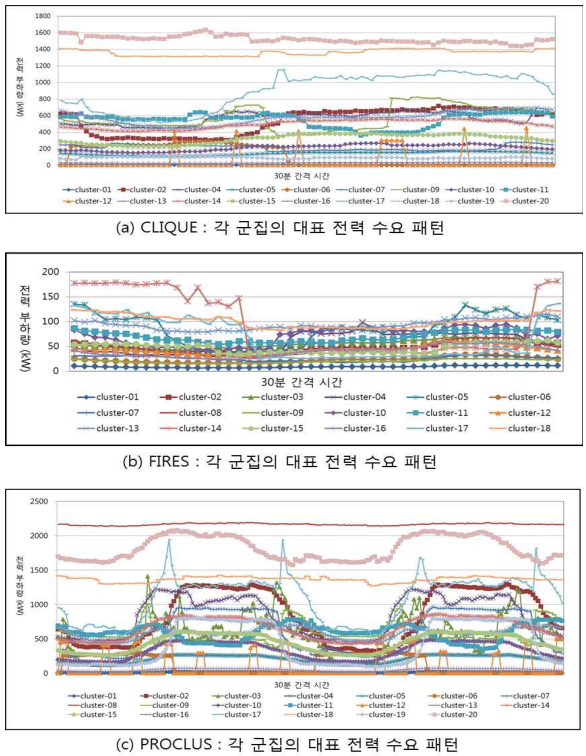
4.2 SMO를 이용한 소지역 단위 전력 수요 패턴 예측

'분류/예측(classification/prediction) 분야에서 사용되는 대부분의 분류 기법들은 훈련데이터의 클래스별 데이터 크기가 동일하다는 전제 하에 디자인 되었기 때문에 클래스별 불균형 데이터 크기 문제에서 자유롭지 못하다. 즉, 다수의 데이터를 보유한 클래스쪽으로 정확도가 높게 나타나고, 소수의 데이터를 보유한 클래스쪽으로 정확도가 낮게 나타나는 경향이 있다. 따라서 각 클래스간 데이터 불균형문제는 분류모델의 성능의 저하를 가져온다.

소지역 전력 수요 예측 문제에서 이러한 클래스 불균형은 예외 일 수 없다. 따라서 이 논문에서 기존의 부분공간 알고리즘에서 적용한 결정트리를 사용을 하지 않는 이유이다.

- SMO(Sequential Minimal Optimization)

부분공간 군집화 결과, 각 군집에 속한 데이터 수는 불균형을 이루며, 이러한 불균형 학습을 위해 클래스에 서로 다른 정규화 값이 부여되는 SVM(Support Vector Machine)의 최적화 문제의 구현에 SMO [15]



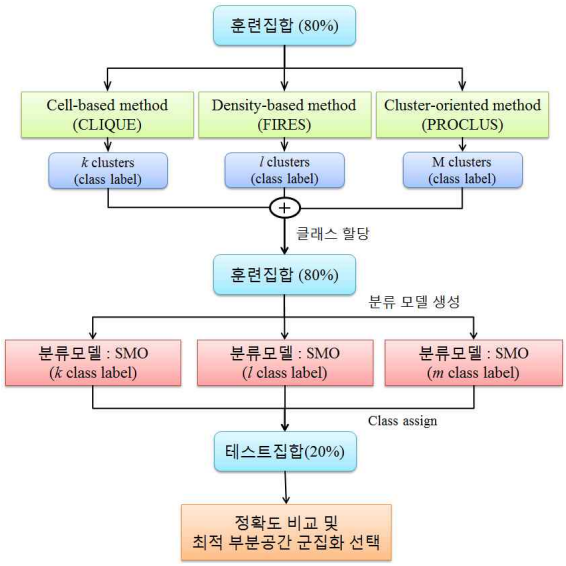
<그림 10> 3가지 부분공간 군집화 수행 결과

알고리즘이 적합하다.

SMO 알고리즘은 반복 시에 선택된 두 라그랑지 변수 α_i와 α_j에 대해 KKT 조건식을 만족하는 해를 찾는 main_routine(), examineExample(i), takeStep(i,j)의 3개의 함수로 구성된다 [16]. main_routine()은 KKT 조건을 만족하는 라그랑지 승수을 학습하기 위해 examineExample(i)를 호출하여 라그랑지 승수를 학습시킨다. examineExample(i)는 α_i와 같이 최적해의 대상인 α_j을 선택하는 함수이다. 선택된 α_i와 α_j 새로운 최적 해를 찾는 과정은 takeStep(i,j)에서 수행된다. SMO는 두 단계의 학습 과정이 반복 수행된다. 먼저, 첫 번째 단계에서 새로운 해의 대상인 2개의 라그랑지 승수 α_i와 α_j를 선택한다. 다음 단계에서는 α_i와 α_j의 새로운 해를 계산한다. 이 두 단계는 모든 라그랑지 승수의 변화가 없을 때 까지 반복되어 종료된다.

4.1절의 부분공간 군집화 문제를 인구통계학 및 지리정보학적 특성 예측과 대표 전력 수요 패턴 예측을 위한 분류 문제로의 전환은 군집화 수행 결과인 군집 라벨을 학습을 위한 클래스 라벨로 간주하며, SMO의 입력 변수인 특징벡터는 부분공간 군집화에 적용된

전체 속성 데이터의 부분 차원을 이용하게 된다. 따라서 3가지 방식의 부분공간 군집화에 사용된 서로 부분 차원들이 SMO 기법의 학습 단계에 사용되며, 서로 다른 분류/예측 결과를 나타나게 된다. <그림 11> SMO를 적용한 분류 기법 적용 단계를 도시화 한 것이며, 평가 내용은 5장에 기술한다.



<그림 11> SMO를 이용한 분류 모델의 생성

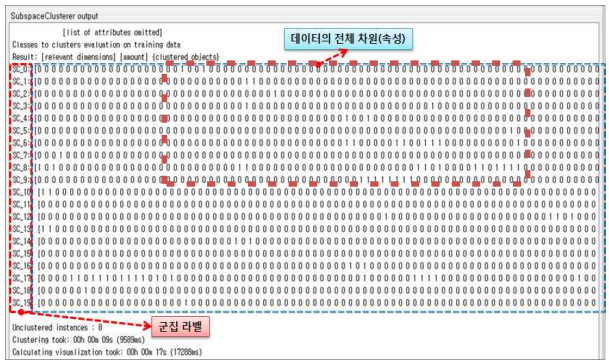
5. 실험 및 평가

소지역 단위 센서스 특성 및 전력 수요 패턴의 군집 분석과 분류/예측을 위해서 서울지역 26개구의 총 16,357개 소지역 센서스 정보와 한전전력연구원에서 제공한 2007년 1월부터 10월까지의 무선부하감시 시스템의 변압기 부하 데이터를 실험에 사용하였다.

5.1 부분공간 군집화 분석

3가지 형태의 부분공간 군집화 수행은 전체 데이터의 모든 차원을 고려한다. 각 군집화 알고리즘 수행 결과 각 군집 구성에 사용된 서로 다른 데이터부분 차원(속성)이 어떤 것인지 확인할 수 있다.

<그림 12>는 clustering-oriented 기법인 PROCLUS 수행 예이며, 프로그램 실행 창에서 'SC0'-'SC20'은 20개의 군집을 나타내는 군집 라벨이며, '0', '1'로 표현된 부분은 전체 입력 데이터에 매칭되는 심볼이다. 각 알고리즘별로 '0'과 '1'의 표현은 부분공간 탐색 방법에 따라 다르며 그 값이 '1'일 경우, 군집 생성 과정에서 '1'에 해당되는 속성 값이 사용되어졌음을 나타낸다. 다시 말해 속성값이 '0'인 경우는, 군집 형성에 전혀 관련 없는 속성으로 무시되었음을 의미한다.



<그림 12> PROCLUS 부분공간 군집화 실행 예

이 논문에서는 군집 알고리즘의 평가를 위해 결정트리를 사용 하지 않았으므로 군집 단계에서의 정확성 평가 내용은 제외된다.

<표 3>은 density-based 방식의 FIRES 알고리즘 실행 결과 요약이다. 총 18개의 군집을 형성하였으며, 센서스 정보 및 1개월 데이터에 대한 30분 간격의 전력 부하량에 대해서 군집 형성 과정에 참여한 정보를 보여준다.

<표 3>의 전력 부하량 분석 과정에서 관련된 시간 차원('1')에서 의미 있는 시간 패턴 추출이 가능하다. 군집 내 특정 시점의 시간(time-stamp) 및 시간 간격(time-interval)을 찾아 낼 수 있다. <표 4>는 <표 3>의 군집 결과로부터 30분 간격 차원을 시간 표현식으로 재구성한 것이다. 월별 전력 부하량이므로 시간 앞에 날짜를 명시한다.

<표 3>의 점선의 직사각형 표시 부분과 <표 4>의 SC6, SC9, SC11, SC13, SC15, SC16, SC17의 시간 정보를 보면 '07.01.01: 01시~03시'는 다수의 군집에 공통적으로 나타난다. 이것은 이 시간대의 전력 부하 패턴이 중요한 요소이며, 따라서 전력분야 전문가 또는 전력 예측시스템 사용자에 의해서 관심 있게 모니터링 해야 하는 시간대임을 암시한다.

<표 3> FIRES 알고리즘 적용 결과 요약

(단, '0'은 생략)

속성		총 18개 군집																
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
센서스 정보	s1	1																	1
	s2	1						1	1				1			1			1
	s3	1							1								1		1
	s4									1									1
	s5		1	1							1			1			1		1
	s6				1		1							1					
	...									...									
	S95				1														
	S96				1							1			1			1	
전력 부하량	0		1																
	1			1															
	2					1	1	1			1		1		1		1	1	1
	3		1					1			1		1		1		1	1	1
	4			1		1	1	1			1		1		1		1	1	1
	5							1			1		1		1		1	1	1
	6							1			1		1		1		1	1	1
	7							1							1		1		
	8	1																	
	9								1										
	10	1																	
	...									...									
	1,440																		

<표 4> 시간 정보를 이용한 시간패턴 분석

군집 라벨	시간 정보
SC0	'07.01.01: 04시, 05시
SC1	'07.01.01: 00시, 01시30분
SC2	'07.01.01: 00시30분, 02시
SC3	-
SC4	'07.01.01: 01시, 02시
SC5	'07.01.01: 01시, 02시
SC6	'07.01.01: 01시 ~ 03시30분
SC7	'07.01.01: 04시30분
SC8	-
SC9	'07.01.01: 01시 ~ 03시
SC10	-
SC11	'07.01.01: 01시 ~ 03시
SC12	-
SC13	'07.01.01: 01시 ~ 03시30분
SC14	-
SC15	'07.01.01: 01시 ~ 03시30분
SC16	'07.01.01: 01시 ~ 03시
SC17	'07.01.01: 01시 ~ 03시

5.2 SMO를 이용한 최적 군집 선정 및 전력 수요 패턴 예측 모델 평가

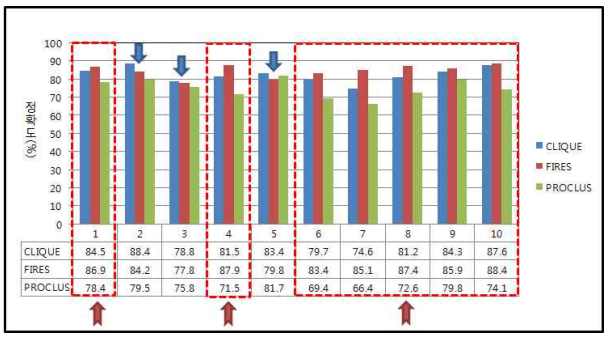
부분공간 군집화 결과 CLIQUE와 PROCLUS의 군집 개수는 20이며, FIRES는 18개의 소지역 전력 부하 패턴 군집을 형성하였다. 소지역 단위 전력 부하 패턴의 최적 군집화 및 전력 수요량 예측을 위해 SMO를 이용하여 분류 모델을 생성한다. 데이터의 80%는 모델 생성을 위한 훈련 데이터로 사용되며, 20%는 모델의 평가를 위한 시험 데이터로 활용한다. 실험에 사용된 전력 부하 데이터는 10개월 분량이다. 국내와 같은 4계절이 뚜렷한 경우에는 최소 몇 년 동안 수집된 데이터를 분석하여야 연 단위의 장기 전력 수요 패턴 예측이 가능하다. 따라서 이 논문에서는 10개월 데이터를 월별로 나눠, 단기 전력 수요 패턴 예측인 월별 분류 모델을 생성하여 실험 평가 한다.

3가지의 부분공간 군집화 결과에 따른 군집 라벨은 클래스 라벨로 간주되고 성능 평가 지표로서 모델의 정확도(accuracy)을 비교한다. 일반적으로 분류 모델의 정확도은 다음 (식 2)와 같이 구한다.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{(식 2)}$$

TP: True Positive, TN: True Negative
FP: False Positive, FN: False Negative

<그림 13>은 1월부터 10월까지 월별 소지역 전력 수요 패턴 분류 결과이다.



<그림 13> 3가지 부분공간 군집 기법에 대한 SMO 정확도 비교

월별 분류 모델의 정확도 비교에서 2월, 3월, 5월을 제외한 나머지에서 density-based 방식의 FIRES가 가장 높은 정확도를 보였다. 더욱이 국내의 특별 전력 수요관리 계절인 여름철을 고려할 때, FIRES 알고리즘의 정확도는 83.4%(6월)에서 88.4%(10월)이다. 따라서 국내 소지역 군집별 전력 수요 패턴 예측에 적합

한 모델은 FIRES임을 알 수 있다.

6. 결 론

국내 전력 수급 위기 상황 대처 및 전력 수요량 예측을 위한 정확한 분석 모델을 위하여 무선부하감시 시스템으로 수집된 변압기 전력 부하 데이터를 통계청 센서스 정보와 결합하여 소지역 단위 전력 수요 패턴 예측을 수행하였다. 센서스 정보 및 30분 단위로 측정된 부하 데이터는 고차원 데이터이므로 효율적인 소지역 군집 구성은 부분공간 군집화 기법을 적용하였으며, SMO 분류/예측 알고리즘을 통해 불균형 데이터 처리 및 정확한 분류 모델을 생성하였다. 실험 결과 센서스 인구통계학 및 지리학적 특성과 전력 수요 패턴에 적합한 부분공간 군집 기법으로는 density-based FIRES 알고리즘이며, 서울지역 대상 총 18개의 소지역 군집을 구성하였다. 또한 FIRES에 의해 생성된 군집라벨에 대한 SMO의 예측 결과는 월별 약 85%의 정확도를 보였다.

이 논문의 한계로는 한전전력연구원에서 제공받은 변압기 데이터가 10개월 이므로 연간 전력 수요 패턴 예측과 같은 장기수요 예측이 불가능하여 단기 예측만을 수행 하였다. 그러나 제안한 부분공간 군집화 및 SMO 기법을 장기간 축적된 데이터에 적용할 경우, 단기 및 중장기 전력 수요 패턴 예측 모델 구축이 가능할 것이다.

참 고 문 헌

[1] 양현석, “예측 불가능한 전력 상황, 수요관리로 완벽 대비한다,” Electric Power, 제72호, pp. 60- 63, 2013.

[2] J. Huang, and R. Shih, "Short-term load forecasting via ARIMA model identification including non-Gaussian process considerations," IEEE Trans. on Power System, vol. 18, pp. 673-679, 2003.

[3] H.G. Lee, B.J. Lee, J.H. Shin, and K.H. Ryu, “Application of Calendar-Based Temporal Classification to Forecast Customer Load Patterns from Load Demand Data,” IEEE Int’l Conf. on Computer and Information Technology, pp. 149-154, 2008.

[4] 박진형, 이헌규, 신진호, 류근호, 김희석, “기대치-최대화 군집 알고리즘과 출현 패턴 마이닝을 이용한 전력 소비 패턴 분석,” 한국정보처리학회, 제15권, 제2호, pp. 261-264, 2008.

[5] 이건학, 최은영, “센서스 GIS와 지리통계시스템의 구축,” 통계연구(2011), 제16권, 제2호, pp. 1-21, 2011.

[6] E. Müller, S. Günnemann, I. Assent, and T. Seidl, “Evaluating Clustering in Subspace Projections of High Dimensional Data,” Int‘l Conf. Very Large DataBases(VLDB), vol. 2, pp.1270-1281, 2009.

[7] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications,” SIGMOD, vol. 27, no. 2, pp. 94-105, 1998.

[8] P. Kriegel, P. Kroger, M. Renz and S. Wurst, “A generic framework for efficient subspace clustering of high-dimensional data,” Int’l Conf. Data Mining, pp. 250-257, 2005.

[9] C. Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park, “Fast algorithms for projected clustering,” ACM SIGMOD, pp. 61-72, 1999.

[10] M. Park, J.Y. Lee, K.H. Ryu, “Subspace Projection based Extraction of Load Shape Factors,” Int’l Conf. Frontiers of information Technology, pp. 12-15, 2012.

[11] S.V. Verdu, “Classification, Filtering, and Identificaton of Eletrical Customer Load Patterns Through the Use of Self-Organizing Maps,” IEEE Trans. Power Systems, vol. 21, pp. 1672-1682, 2006.

[12] E. Müller, S. Günnemann, I. Assent, and T. Seidl, “Evaluating Clustering in Subspace Projections of High Dimensional Data, http://dme.rwth-aachen.de/OpenSubspace/,” Int’l Conf. Very Large Data Bases(VLDB), Lyon, France, 2009.

[13] P. Saeed, and A.N. Somaye, "A New Semantic Kernel Function for Online Anomaly Detection of Software" ETRI Journal, vol. 34, pp. 288-291, 2012.
[14] W. Jun, and M-Y. Lu, "Asymmetric Semi-Supervised Boosting Scheme for Interactive Image Retrieval," ETRI Journal, vol. 32, pp. 766-773, 2010.
[15] C. Lee, and M-G, Jang, "A Prior Model of Structural SVMs for Domain Adaptation," ETRI Journal, vol. 33, pp. 712-719, 2011.
[16] 김광성, 황두성, "불균형 데이터 학습을 위한 지지벡터기계 알고리즘," 한국컴퓨터정보학회지 제15권, 제7호, pp. 11-17, 2010.

이 헌 규 (Heon Gyu Lee)

- 정회원
- 충북대학교 전자계산학과 이학석사
- 충북대학교 전자계산학과 공학박사
- 한국전자통신연구원 융합기술연구부문 선임연구원
- 관심분야 : 데이터마이닝, 기계학습, 패턴인식, 바이오인포매틱스, 바이오매디컬, 데이터베이스, GIS 등

신 용 호 (Yong Ho Shin)

- 정회원
- 서울대학교 산업공학과 공학학사
- 한국과학기술원 산업 및 시스템공학과 공학석사
- 한국과학기술원 산업 및 시스템공학과 공학 박사
- 영남대학교 경영학부 조교수
- 관심분야 : 경영과학, 데이터마이닝, Machine Learning, 정보시스템, Formal Model 등

논 문 접 수 일 : 2013년 04월 26일
1차수정완료일 : 2013년 05월 09일
2차수정완료일 : 2013년 06월 07일
게 재 확 정 일 : 2013년 06월 21일