Journal of the Korean Data & Information Science Society
2015, **26**(5), 1105–1115

http://dx.doi.org/10.7465/jkdi.2015.26.5.1105
한국데이터정보과학회지

자연재해 분석을 위한 빅데이터 마이닝 기술[†]

김영민[1] · 황미녕[2] · 김태홍[3] · 정창후[4] · 정도헌[5]

[12345]한국과학기술정보연구원

접수 2015년 8월 5일, 수정 2015년 9월 7일, 게재확정 2015년 9월 16일

요 약

자연재해 빅데이터 분석은 현재 소셜 미디어 데이터 등 텍스트 데이터를 중심으로 시작되고 있으며 이는 재난관리의 네 단계인 예방, 대비, 대응, 복구에서 마지막 두 단계에 주로 해당된다. 반면 기상 데이터 자체에 대한 빅데이터 분석은 사전 관리에 해당하는 예방, 대비 단계에 활용될 수 있어 이와 관련한 연구 사례에 대한 체계적인 정리가 필요하다. 본 논문은 리뷰 논문으로서, 자연재해 영역에서 텍스트 데이터 외의 빅데이터를 다루는 분석 기술들에 대해 소개한다. 이를 위해 기상 관련 분야에서 사용되고 있는 데이터 마이닝 및 기계 학습 기술들을 살피고 각 기상 데이터의 특성에 맞춰 기존의 기술들이 어떻게 변형되는 지 밝힌다. 우선 2절에서 빅데이터, 데이터 마이닝, 기계 학습에 대한 기본 개념을 설명하고 3절에서 데이터 마이닝 및 기계 학습 기술의 실제 적용 사례를 상세히 정리한다. 4절에서는 자연재해 대응에 이러한 기술들이 직접 활용되는 예를 소개하고 마지막에 결론으로 마무리한다.

주요용어: 기계 학습, 기상 데이터, 데이터 마이닝, 빅데이터, 자연재해.

1. 머리말

불과 십여 년 전부터 등장하기 시작한 빅데이터는 큰 관심을 받으며 학계를 넘어 일반인들에게까지 친숙한 개념으로 자리 잡아 가고 있다 (Magoulas와 Lorica, 2009). 이러한 급속한 유행으로 인하여 빅데이터 분석에 사용되는 데이터 마이닝이나 기계 학습의 중요성이 부각되었고, 전통적으로 그러한 기술에 그리 관심을 두지 않았던 다양한 영역들에서 빅데이터를 활용할 방법을 적극적으로 찾게 되었다. 그리고 그 대표적인 분야 중 하나가 재난재해 분야이다.

지난 2013년 5월 미국 국립과학재단 (NSF)과 일본 과학기술진흥기구 (JST)가 주최한 '빅데이터와 재난관리' 워크샵에서는 이 분야의 다양한 연구자들이 모여 관련 연구 현황에 대한 논의를 하고 리포트를 작성했다 (http://grait-dm.gatech.edu/wp-content/uploads/2014/03/BigDataAndDisaster-v34.pdf). 여기서 정의하는 재난 분야의 빅데이터 종류는 크게 두 가지로, 환경 센서 데이터와 소셜 미디어 데이터로 나뉘는데, 재난관리의 네 단계인 예방 (prevention), 대비 (preparedness), 대응 (response), 복구 (recovery)에 모두 빅데이터를 활용하고 있다. 일본 Tohoku 지진에서의 센서 데이터를

[†] 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 SW컴퓨팅산업융합원천기술개발사업의 일환으로 수행하였음 (B010-15-0353, 빅데이터 통합 모니터링 및 분석을 위한 고성능 데이터베이스 솔루션 개발).
[1] (305-806) 대전광역시 유성구 대학로 245, 한국과학기술정보연구원, 선임연구원.
[2] (305-806) 대전광역시 유성구 대학로 245, 한국과학기술정보연구원, 선임연구원.
[3] (305-806) 대전광역시 유성구 대학로 245, 한국과학기술정보연구원, 선임연구원.
[4] (305-806) 대전광역시 유성구 대학로 245, 한국과학기술정보연구원, 선임연구원.
[5] 교신저자: (305-806) 대전광역시 유성구 대학로 245, 한국과학기술정보연구원, 선임연구원.
E-mail: heon@kisti.re.kr

통한 지진 감지와 소셜 네트워크를 통한 현황 파악 및 복구 지원, 2015년 네팔 지진에서의 위성 데이터 패턴 분석을 통한 지진 지역 파악 등이 그 예이다. 한편 2015년 일본 센다이시에서 개최된 유엔 방재 세계 회의 (Third UN World Conference on Disaster Risk Reduction)에서 발표된 바로는 최근 10년간 일어났던 재난재해의 87%가 기상 관련 자연재해였다. 여기서 정의하는 재난 분야의 빅데이터 종류는 크게 두 가지로, 환경 센서 데이터와 소셜 미디어 데이터로 나뉘는데, 재난관리의 네 단계인 예방 (prevention), 대비 (preparedness), 대응 (response), 복구 (recovery)에 모두 빅데이터를 활용하고 있다. 일본 Tohoku 지진에서의 센서 데이터를 통한 지진 감지와 소셜 네트워크를 통한 현황 파악 및 복구 지원, 2015년 네팔 지진에서의 위성 데이터 패턴 분석을 통한 지진 지역 파악 등이 그 예이다. 한편 2015년 일본 센다이시에서 개최된 유엔 방재 세계 회의 (Third UN World Conference on Disaster Risk Reduction)에서 발표된 바로는 최근 10년간 일어났던 재난재해의 87%가 기상 관련 자연재해였다 (http://www.unisdr.org/archive/42862). 그러나 재난 관련 빅데이터 분석의 상당수가 기상 데이터 자체보다는 재난에서 파생된 텍스트 기반의 소셜 데이터 분석에 집중하고 있다. 소셜 데이터가 재난이 발생한 이후의 대응, 복구 단계에서 많이 활용되는 반면에, 기상 관련 데이터는 재난이 발생하기 이전의 예방, 준비 단계에 유용하게 활용될 수 있기 때문에 관련 재해의 피해를 최소화하기 위해서는 기상 관련 데이터 자체에 대한 심층 분석이 필수불가결하게 요구되고 있는 상황이다.

본 논문에서는 자연재해 영역에서 텍스트 데이터 외의 빅데이터 분석이 어떻게 이루어지고 있는지 살펴보기 위해 기상관련 분야에서 활용되고 있는 데이터 마이닝과 기계 학습 기술들을 리뷰하고자 한다. 즉, 빅데이터 분석의 두 가지 키워드인 대규모 데이터와 분석 기술 중 후자에 집중한다. 빅데이터가 이슈화 되면서 데이터의 규모에만 관심이 집중된 반면 데이터에서 의미를 찾아내는 도구인 분석 기술에 대해서는 충분한 논의가 부족했던 것이 사실이다. 따라서 기상 분야에서의 마이닝 사례를 정리하는 것이 현재 시작되고 있는 자연재해 빅데이터 분석 연구에 도움이 되리라 생각한다.

본 논문의 구성은 다음과 같다. 우선 2절에서 빅데이터, 데이터 마이닝, 기계 학습에 대한 기본 개념과 그들의 관계를 정리하고 3절에서는 기상관련 분야에서 활용되고 있는 데이터 마이닝과 기계학습 기술들을 소개한다. 4절에서는 자연재해 대응에 직접 활용되는 기술의 예를 들어 설명하고 마지막 5절에서 결론은 맺는다.

2. 빅데이터 마이닝과 기계 학습

본 절에서는 핵심 키워드인 빅데이터, 데이터 마이닝, 기계 학습에 대해 간략하게 소개한다.

2.1. 빅데이터

수년 전부터 데이터 분석 분야를 대표하는 하나의 현상을 꼽는다면 단연 빅데이터 일 것이다. 빅데이터란 단순히 대용량 데이터를 처리하고 분석하는 것을 넘어 기하급수적으로 증가하는 데이터에 효과적으로 대응하기 위한 방법론들을 총칭하는 용어이다. 기술적으로는 크게 두 가지 관점으로 나눠볼 수가 있는데, 대용량 데이터를 효율적으로 처리할 수 있는 분산 처리 기술의 관점과 데이터를 효과적으로 분석할 수 있는 방법론적인 관점이다. 이 두 가지 모두 빅데이터 분석에 있어서 중요한 요소이지만 한국 빅데이터 연구의 초창기에는전자를 위한 시스템 구축에 치중했다고 볼 수 있다. 반면 분석을 위한 방법론들은 기존의 단순한 통계적 분석 틀을 벗어나지 못하는 경우도 많았으며 이를 넘어 보다 지능적인 분석은 최근에야 시도되고 있는 상황이다 (Kim, 2013). 빅데이터 분석 대상은 텍스트, 이미지, 음성, 영상, 시그널, 구조화된 정형 데이터 등 데이터의 종류를 가리지 않지만 주로 텍스트 데이터를 다루는 연구가 많이 이루어졌다. 실제 적용에 있어서도 소셜 미디어 데이터 분석을 통한 마케팅 활용, 재난재해 이벤트 감지 등 텍스트 데이터를 활용하는 사례가 많다. 하지만 멀티 소스로 주어지는 데이터가 증가함

에 따라 이미지와 텍스트, 정형과 비정형 데이터 등 이종의 데이터를 통합한 분석의 필요성이 점점 더 증가하고 있는 상황이다.

2.2. 데이터 마이닝과 기계 학습

빅데이터 분석을 위해 완전히 새로운 기법을 개발하는 경우는 드물다. 대부분 기존 데이터 분석에 사용하는 데이터 마이닝 및 기계 학습 기법을 활용하며 빅데이터 환경에 맞도록 입력 데이터를 분산 처리하거나 알고리즘을 변형시킨다. 데이터 마이닝은 주어진 데이터를 자동으로 처리·분석하여 의미 있는 정보를 찾아내는 방법론들을 일컫는 학문 분야로서, 각종 통계적 분석 및 기계 학습 기술들이 사용된다. 데이터 마이닝이 의미 있는 정보를 찾아낸다는 목적에 초점을 둔 개념이라면, 기계 학습은 실제 사용되는 기술에 초점을 둔다.

기계 학습은 인공지능의 한 분야로서 기계가 주어진 데이터로부터 자동으로 학습할 수 있도록 하는 수리적 모델과 알고리즘을 연구하고 개발하는 학문 분야이다. 기계라 함은 컴퓨터와 같이 반복적인 수치 계산이 가능한 도구를 가리키며 협의의 개념으로는 기계 학습 기술이 적용된 소프트웨어라고 볼 수도 있다. 여기서 학습은 인간의 지적 활동을 시뮬레이션 하는 등의 고차원적인 것을 뜻하지는 않는다. 대신 주어진 데이터를 사용하여 자동 인식이나 분류 등을 수행하기 위해 수리적 모델을 만들어 통계적, 확률적 방법을 통해 반복적으로 계산하는 과정을 학습이라 일컫는다. 이러한 반복적인 계산 방법은 수학적으로 검증된 공식을 따르며 그 계산 과정은 수식으로 표현된 모델과 구별되어 알고리즘이라고 불린다.

데이터 마이닝과 기계 학습은 이렇게 정의하는 수준이 다르기 때문에 이들의 관계가 포함관계로 정의될 수는 없으나 간혹 같은 것을 지칭하는 개념으로 혼용되기도 한다. 정리하자면, “데이터 마이닝에 기계 학습 기술을 사용할 수 있다” 정도가 이들의 관계를 올바르게 표현하는 방식 중 하나이다. 본 논문에서는 데이터 마이닝에 사용되는 기술 중 특히 기계 학습 기술을 다루므로 기술적인 방법론을 지칭할 시 데이터 마이닝 대신 기계 학습이라는 용어를 사용하도록 한다.

3. 기상 데이터 마이닝

본 절에서는 기상 데이터 처리에 기계 학습 기법들이 어떻게 적용되는지 살펴본다. 우선 시계열로 표현되는 다양한 기상 데이터 및 그것들에게 공통으로 적용될 수 있는 기술들을 소개한다. 또한 크게 분석을 예측과 패턴 인식으로 나누어 각 분야의 연구 현황을 밝힌다.

3.1. 시계열 기상 데이터

기온, 상대 습도, 바람 속도, 바람 방향, 대기 압력, 태양 복사, 오존 등 대부분의 기상 데이터는 시계열로 표시된다. 이러한 시계열 데이터는 전통적으로 확률 과정론적인 방법을 사용하여 분석하는데, 크게 자기회귀 (AutoRegressive; AR) 모델, 이동평균 (Moving Average; MA) 모델, 자기회귀 이동평균 (AR moving average: ARMA) 모델, 자기회귀 직접 이동평균 모델 (AR integrated moving average; ARIMA), 그리고 마지막으로 마코프 연쇄 (Markov Chain) 모델 카테고리로 나뉜다 (Mellit 등, 2012). 시계열적인 특성을 확률 과정으로 직접 모델링했기 때문에 논리적으로 설득력이 있지만 시계열 특성이 어떠한 패턴으로 수렴하지 않는다면 성능이 좋지 않을 수 있다는 단점이 있다.

90년대 말 경부터 기상데이터 분석에 시계열을 직접 모델링 하지 않는 기계 학습 기법들이 활발히 사용되기 시작했다. 인공 신경망 (artificial neural network; ANN)과 서포트 벡터 머신 (support vector machine; SVM) (Cortes와 Vapnik, 1995; Choi 등, 2013)이 가장 많이 쓰이고 있는 대표적인 기술인데 후자의 경우 본래 분류 목적으로 주로 쓰이나 여기서는 기상 예측을 위해 회귀

분석용으로 변형한 서포트 벡터 회귀 (support vector regression; SVR) (Drucker 등, 1997)을 사용한다. 이러한 기술을 적용하기 위해 기상 시계열 데이터는 다음과 같이 입력데이터와 예측을 위한 출력데이터로 정의된다. 일단 t라는 시각에서의 어떤 기상 데이터 예측값을 x^t라 했을 때, 이 값은 그 이전의 기상 관측데이터 집합인 $\{x^{t-1}, x^{t-2}, \cdots, x^{t-k+1}, x^{t-k}\}$을 사용하여 예측한다. 이 시계열 데이터를 입력 벡터, $\left(x^{t-1}, x^{t-2}, \cdots, x^{t-k+1}, x^{t-k}\right)$라고 가정하고 이 데이터를 입력값으로 하는 모델을 학습하여 x^t를 예측하는 것이 목적이다. 모델은 다음과 같이 회귀 함수 형태로 표현된다: $x^t = f\left(x^{t-1}, x^{t-2}, \cdots, x^{t-k+1}, x^{t-k}\right)$.

3.2. 예측

기상 데이터를 기계 학습으로 분석한다고 했을 때 가장 쉽게 생각할 수 있는 것이 과거 관측 데이터를 기반으로 기온, 바람 등 각 기상 요소에 대한 예측을 하는 것이다. 그러나 기상학에서 일기 예보용으로 사용하고 있는 수치 예측 모델과는 달리 주로 각 기상 요소에 대한 단기 예측용으로 쓰이기 때문에 기계 학습이 수치 모델을 대체할 수 있다는 것은 아니다. 본 절에서는 앞서 언급한 ANN과 SVM을 중심으로 기계 학습 기법들이 기상 데이터 예측에 어떻게 활용되고 있는지 최근 연구 사례를 통해 살펴본다.

3.2.1. 기온

ANN은 회귀분석을 대신하여 최초로 시계열 기상 데이터에 적용된 기계 학습 기법으로, 기온 단기 예측에서 기존의 통계적인 접근 방법에 비해 좋은 성과를 보였다. 초창기 연구로는 Snell 등 (2000)이 11개 특정 위치의 지표면 기온을 예측하기 위해 ANN을 사용한 사례가 있다. 주변 지역의 기온 데이터를 이용하여 목표 지역의 기온을 예측하는 내삽에 ANN을 사용했던 최초의 연구 중 하나이다. 기존의 마이닝 기술들과 (spatial average, nearest neighbor, inverse distance methods 등) 비교한 실험 중 94%의 실험 케이스에서 좋은 결과를 보였다. Tasadduq 등 (2002)은 ANN을 사용하여 시간대 별 평균 온도를 24시간 미리 예측하는 연구를 했다. 사우디 아라비아 제다 지역 연안의 1년 동안의 온도 데이터를 이용하여 모델을 학습했다. 삼년 동안의 데이터에 대해 실측값과 예측값의 차이를 비교하였는데, 각 연도별 데이터에 대해 3.16%, 4.17%, 2.83%의 평균 편차를 보였다. ANN이 시계열 분석에서 많은 성과를 보였지만 수학적으로 최적화된 해를 보장하는 것은 아니며 지역 해 (local minimum)나 과적응 (overfitting) 문제 등에 취약하다. SVR은 이러한 약점을 보안하며 ANN을 대체하였는데, Radhika와 Shashi (2009)는 SVR을 사용하여 일일 최고 기온을 예측해냈다. 입력데이터로는 예측 날짜 이전의 n일에 대한 최고 기온을 사용했으며 n 이라는 값은 실험을 통해 유추해냈는데 결과적으로 기존 ANN과 비교하여 좋은 성능을 보였다.

3.2.2. 태양 에너지

재생 에너지 활용에 대한 관심이 높아지면서 태양 에너지를 효율적으로 활용하기 위해 일사 강도를 정확히 예측하는 연구가 최근 많이 시도되고 있다. Voyant 등 (2011)은 일일 일사량 (daily global radiation) 예측에 ANN을 사용하였는데, 주목할 만한 점은 입력값으로 일사량 과거 데이터뿐만 아니라 다른 기상 과거 데이터도 활용한다는 것이다. 이 두 가지 종류를 내부 (endogenous), 외부 (exogenous) 데이터라 명명하고 ARIMA 기법과 ANN을 사용하여 외부데이터의 유효성을 검토하였다. 프랑스 코르시카 섬의 두 개의 관측지점에서의 2006년에서 2007년 사이의 데이터를 활용하여 실험했으며 ANN이 ARIMA의 성능을 능가함을 밝혀냈다. 날씨가 좋은 날이 많은 여름에는 내부데이터를 쓰는 것으로 충분했으나 흐린 날이 많은 겨울에는 외부데이터까지 활용하는 것이 더 정확한 예측 결과를 보여줬다. 같은 해 Sharma 등 (2011)에 의해 비슷한 연구가 진행되었는데, 입력값에서 다른 점은 외부 데이터로 미

국 National Weather Service (NWS)의 기상 예측 데이터를 사용한다는 것이다. 한 시간 전 예측을 했던 Voyant 등의 연구와는 달리 세 시간 전에 예측하는 것을 목표로 모델링 했으며 선형 최소 제곱 회귀 (linear least squares regression)과 SVR을 사용하여 실험했다. 2010년 10개월 동안 특정 기상관측소에서 수집된 데이터를 활용하였으며, 8개월 동안의 데이터로 학습을 하고 나머지로 테스트를 했다. SVR의 경우 다양한 내부 커널을 테스트 했는데, RBF (radial basis function) 커널이 가장 좋은 결과를 내었다. 최종적으로 기존의 예측 모델에 비해 27% 더 정확함을 보여주었다. 한편 비슷한 종류의 데이터를 이용하지만 전혀 다른 접근 방법을 사용한 연구가 그 다음해 이루어졌다 (Chakraborty 등, 2012). Chakraborty 등은 태양열 에너지 (PV) 생산량을 예측하기 위해 세 가지 서로 다른 예측 모델을 통합하는 베이지안 앙상블 기법을 제안했다. 두 가지 모델은 전통적인 접근 방법을 쓰는데, 기상 예측 데이터를 활용하여 PV를 예측하는 Naive Bayes 모델과 시계열로 주어지는 당일의 PV 값을 활용한 KNN (K-nearest neighbor) 기반 예측이 그것이다. 반면 과거 PV 값을 사용하는 세 번째 모델은 시계열 데이터에 숨겨진 패턴을 잡아내기 위해 이 과거 데이터를 심볼로 패턴화한 스트림 값으로 변환한다. 이어 가장 자주 등장하는 패턴을 찾아 PV를 예측한다. 세 가지 예측 모델을 통합하는 앙상블 모델을 마지막에 적용하여 최적의 값을 예측했으며 Sharma 등의 연구보다 좋은 성능을 보였다.

3.2.3. 바람 및 풍력 발전

기상 요소 중 바람과 관련해서는 크게 방향 예측과 속도 예측이 있다. 풍력 발전기 운영 계획 및 실행을 위해서는 단기 예측 (1시간∼72시간)이 중요하며 확률, 통계기반의 연구들도 활발히 이루어져왔다. Foley 등 (2012)은 최근 풍력 발전 예측 분야에 대한 리뷰 논문을 작성했는데, 크게 두 가지 그룹으로 나누어 지금까지의 연구를 정리하였다. 첫 번째 그룹은 기계 학습 등 통계 기반의 방법론이고 두 번째 그룹은 수치 기상 모델을 사용하는 예측이다. 대부분의 경우 전자의 정확도가 높으나 일일 및 시간별 예측에서는 후자의 역할도 중요해진다. 통계적 방법론들이 정확도가 높았던 관계로 많은 연구가 이루어졌으며 시계열 데이터에 사용되는 대부분의 방법론들이 활용되었다. ANN을 비롯, 퍼지 시스템, 유전자 알고리즘, SVR, 베이지안 방법론 등이 사용되었으며 수치 모델과 혼합된 모델들도 개발되었다. 그 중 몇 가지를 들자면, Mohandes 등 (2004)이 초창기에 SVR을 이용하여 바람의 속도를 측정했고, Jursa와 Rohrig (2008)가 ANN과 KNN을 사용한 최적화 모델을 개발했다. He 등 (2011)은 바람 프로파일을 7개의 자질로 정의하고 이 중 체감온도 일부를 삭제한 후, 불완전한 데이터로 ANN과 앙상블 학습을 이용해서 바람 속도를 예측하는 색다른 연구를 수행했다. Ohashi와 Torgo (2012)은 시계열 정보뿐만 아니라 공간정보도 활용한 시공간 데이터에 다양한 기계학습 방법들을 적용하여 기존 연구들보다 좋은 성능을 보여주었다.

3.2.4. 강우 및 하천유량

데이터 마이닝을 이용하여 물을 분석한 연구로는 크게 강우량 예측과 하천유량 예측으로 나뉜다. Kusiak 등 (2012)은 레이더 반사율을 이용한 강우량 예측에 마이닝 기법을 적용했는데 ANN, random forest, 분류 및 회귀 분석, SVM, KNN 등 다섯 가지 기계학습 알고리즘을 비교하였다. 입력값으로 레이더 반사율과 Tipping bucket (TB)을 사용하였으며 ANN이 가장 좋은 성과를 보였다. Rasouli 등 (2012)은 캐나다 한 지역의 하천유량 예측에 다음과 같은 다섯 가지 기계학습 방법을 사용하였다: Bayesian neural network (BNN), SVR, Gaussain Process (GP), multiple linear regression (MLR). 기상-수문 관측 데이터를 기본으로, GFS 모델의 기상 예측 데이터와 기후 지수까지 입력데이터로 사용하여 최종적으로 MLR이 가장 좋은 성능을 보임을 밝혔다. 짧은 기간일수록 GFS 데이터, 반대의 경우 기후 지수가 의미가 있었다. 이 연구가 1∼7일의 선행 시간이라는 단기 예측을 목표로 했다면 Kalra

등 (2013)은 1년이라는 장기간의 선행 시간에 대한 하천유량 예측을 수행하였다. 미 서부와 같은 물 부족 지역에서의 수자원 관리와 운용을 위해서는 이러한 장기간의 예측이 필요한데 SVR 기반의 방법이 ANN이나 MLR 보다 정확히 예측함을 보였다. 하천유량 예측에서 나아가 가뭄에 대비한 물 배분 정책에도 마이닝적 기법을 활용할 수 있다. Chang 등 (2013)은 시스템 분석적인 방법과 ANN의 일종인 Adaptive Neuro-Fuzzy Inference System (ANFIS)를 결합한 기법을 대만 Shihmen 저수지의 물 배분 정책에 활용하였다. 입력값으로는 물세 할인율 (water discount rate), 수문학적 상황에 따른 저수량 초과 확률, 물 유입량이 있으며 예측해야할 값은 물 부족 수준 (water deficiency level)이다. 이와 같은 방법을 사용하여 가뭄에 대비한 최적의 물세 할인율 및 관개시설 운영 정책을 제시했다.

3.3. 패턴 인식

기상 데이터 분석에서 예측 이외에 마이닝적 접근 방법이 활발하게 사용되고 있는 분야는 패턴 인식이다. 주로 이미지 데이터 분석에 쓰이며 지표면의 상태 인식 및 분류 등에 활용된다. 동일한 방법이 기상 데이터 뿐 아니라 자연재해와 직접적으로 연관된 지형 분석에서도 쓰이므로 본 절에서는 지형 데이터까지로 범위를 확장한다.

3.3.1. 지표면 분석

지형 데이터 패턴 분석에는 주로 위성 데이터를 사용한 이미지 분류 기법을 사용한다. 위성 데이터를 사용할 수 있게 된 이래로 토지 피복이나 이용을 자동으로 인식하는 연구가 지금까지 다양하게 이루어져왔다. 최근 들어 리모트 센싱 데이터 (RSI) 정보가 급격히 늘어나면서 기존의 이미지 분류에 RSI를 결합한 분석이 많이 이루어지고 있다 (Dos Santos 등, 2012). 이러한 분석은 전통적인 분류 기법을 활용하는 것이므로 현재까지 개발된 대부분의 분류 모델을 적용할 수 있다. Vatsavai 등 (2011)은 고화질의 RSI 데이터를 분석하기 위해 10가지의 서로 다른 기계 학습 모델을 테스트했는데, 최대 우도 (maximum likelihood)나 로지스틱 회귀 (logistic regression) 같은 간단한 방법들이 오히려 좋은 결과를 낸다고 밝혔다. Petropoulos 등 (2012)은 토지피복과 토지 이용 매핑을 위해 초분광 영상을 입력데이터로 사용하여 SVM과 ANN을 적용한 연구를 했으며 기존의 이미지 데이터를 사용한 것보다 좋은 성능을 보여줬다.

3.3.2. 해수 해양

해양과 관련해서는 해양과 대기가 대지 기후에 미치는 영향력을 분석하기 위해 마이닝 기법들이 사용되기도 한다. Storch 등 (1999)이 PCA 와 SVD를 활용하여 기후 지수 (climate index)를 찾는 방법을 개발한 이래 이러한 차원 축소 기법들은 기후 지수 분석에서 주요한 방법론으로 자리 잡았다. 이후 Steinbach 등 (2003)은 클러스터링 기법을 활용하여 sea surface temperature (SST)와 sea level pressure (SLP)에 적용함으로써 PCA 등을 대체했고 이후로도 기후 지수를 찾는데 클러스터링 기법들이 많이 사용되었다 (Race 등, 2010). 한편 클러스터의 개수를 미리 임의로 지정해야 하는 전통적인 클러스터링 기법의 문제가 기후 지수 분석에서도 주요 쟁점으로 등장하게 되었다. 이에 대한 대안으로 complex network 를 사용한 방법론들이 다양하게 연구되었다 (Tsonis 등, 2006; Donges 등, 2009; Steinhaeuser 등, 2010; Steinhaeuser 등, 2011).

3.3.3. 환경

이번 절에서는 위의 카테고리들과 중첩될 수는 있지만 '환경'이라는 키워드로 묶일 수 있는 다른 두 가지 연구에 대해 짧게 소개한다. Peters 등 (2009)은 random forest라는 앙상블 러닝 기법을 사용하

여 식물 분포를 모델링하는 방법을 제안했다. 여기서 특이할만한 점은 불확실성이라는 요소를 평가하여 모델링했다는 것인데, 환경 변수를 공간 내삽할 때와 식물 종류 클러스터링에서의 불확실성을 각각 Gaussian simulation과 pseudo-randomization 테스트를 통해 평가했다. Vincenzia 등 (2011)은 해양 동물 분포 추정에 마이닝 기법을 활용했는데, 이탈리아 베니스 석호 지역의 바지락 분포를 예측하는 모델을 개발했다. 역시 random forest를 활용했으며 연구 결과, 바지락 분포에 가장 많이 영향을 미치는 요소는 침전물 중 모래 비율, 염분, 해수 깊이 이렇게 세 가지로 판별되었다.

4. 자연재해 데이터 분석

3절에서 자연재해 분석의 기반이 될 수 있는 데이터의 종류 및 분석 목적별로 데이터 마이닝 적용 사례를 리뷰했다면 본 장에서는 자연재해에 직접적으로 사용된 마이닝 기법들에 대해 소개한다. 자연재해 중 지진, 산불, 폭풍우, 유해적조에 대한 최근의 연구 사례를 리뷰한다.

4.1. 지진 예측

지진 예측은 어려운 문제이나 발생 시 큰 인명·재산 피해를 내는 위협적인 자연재해 중 하나이며 지금까지 꾸준히 연구되어 왔었다 (Panakkat와 Adeli, 2008). 전통적으로 지구물리학 분야에서의 지진 예측 연구는 대부분 지진 발생 순환 주기를 기반으로 한다. 한편 최근 들어 데이터 기반의 기계 학습 방법론들을 적용한 연구가 많이 진행되었는데 대부분 ANN을 사용하고 있다. 입력 데이터 및 ANN 모델의 세부 종류, 학습 방법들이 변경되며 다양한 연구들이 진행되고 있으며 특히 지진의 규모 예측에서 좋은 성과를 보이고 있다. Kulahci 등 (2009)은 라돈 가스 등을 입력값으로, Moustra 등 (2011)은 지진 전기적 신호를 입력값으로 사용한다. 최근에는 radial basis function 기반의 ANN을 적용한 지진 예측 연구가 이루어졌는데 (Alexandridis 등, 2014) 소규모의 데이터에도 좋은 성능을 보이기 때문에 데이터 수집이 어려운 지진 예측에 적합한 모델이다.

4.2. 산불 예측 및 감지

산불과 관련해서 마이닝 기법을 적용한 연구는 위성 자료 패턴 인식 등 90년대 중반부터 꾸준히 있어 왔으며 (Vega-Garcia 등, 1996; Arrue 등, 2000) 최근 들어 센서 데이터 활용이 높아지면서 실시간 산불 예측에 기상 센서 데이터가 활용되기 시작하였다. 그 최초의 연구 중 하나가 Cortez와 Morais (2007)의 논문으로, 산불 발생과 그 규모를 예측하는 데 SVM과 random forest 등 다섯 가지 방법을 테스트 하였다. 센서로부터 수집된 기상 데이터 중 온도, 상대습도, 비, 바람이 산불에 영향을 미치는 주요 인자로 밝혀졌으며 SVM이 좋은 성능을 보였다. Sakr 등 (2010)도 기상 데이터를 이용한 유사한 연구를 진행하였는데, 특이할 만한 점은 1에서 4 사이의 산불 위험 지수를 도입하고 이를 예측하는 방법을 개발했다는 것이다. SVM을 기반으로 한 예측 알고리즘을 제안했는데, 3개의 SVM 모델을 조합하여 산불 위험 지수를 예측해내는 구조로 이루어져 있다. 2000년에서 2008년까지 6월에서 10월 사이에 레바논에서 일어난 산불 데이터로 검증했으며 평균 88%의 예측 정확도를 보였다. 한편 Angayarkkani와 Radhakrishnan (2010)은 ANN을 이용하여 공간 이미지 데이터로부터 산불을 감지해내고 산불이 일어난 지역을 자동으로 추출하는 연구를 수행했다. 그리고 실제 센서로부터 수집된 이미지 데이터를 활용하여 모델을 검증하였다.

4.3. 기타

지진과 산불 외에도 다양한 자연재해에 마이닝 기법들이 활용되고 있다. Li 등 (2008)은 메조사이클론 (직경 16km까지의 작은 사이클론) 감지에 ANN과 클러스터링, 이미지 처리 등을 사용하였는데, 입력데이터로는 대규모 실시간 기상 데이터와 시뮬레이션 자료를 사용하였다. 미국 NSF의 지원을 받아 10개의 연구 기관이 참여한 프로젝트인 Linked Environments for Atmospheric Discovery (LEAD)의 일부로, 중간 규모의 기상 관련 실시간 환경에 대응할 수 있는 프레임워크를 만드는 연구를 소개한다. 한편, 유해적조 등 해양생태계 오염도 중요한 자연재해 중 하나이다. Muttil와 Chau(2007)은 ANN과 유전 프로그래밍 (GP)을 사용하여 유해적조 이동 모델링에서의 주요 입력 변수를 선택하는 연구를 수행했다. 이 연구는 홍콩 토로항의 수질 측정에 활용되어 실제 입력 변수 선택에 유효한 방법임을 밝혀냈다.

5. 결 론

자연재해 빅데이터 분석은 이제 시작 단계에 있으며 앞으로 이를 효과적으로 수행하기 위해서는 현재 사용되고 있는 기술 현황을 제대로 파악하는 것이 중요하다. 하지만 현재까지 빅데이터 분야는 대규모 데이터를 효율적으로 분산처리 하는 것에 집중되어 왔으며 그 대상은 주로 비정형 텍스트 데이터이다. 자연재해 분야 역시 분석적인 측면에서는 아직 소셜 미디어 데이터 분석이 주를 차지하고 있다.

본 논문은 기상 데이터에 적용된 데이터 마이닝 및 기계 학습 기술들의 현황을 정리함으로써 자연재해 분야의 빅데이터 분석이 나아가야 할 방향에 대한 레퍼런스를 제공한다. 현재 시계열 기상 요소 예측에서는 주로 SVM과 ANN이 사용되고 있으며, 전통적인 기계 학습 기법인 퍼지 시스템, 유전자 알고리즘, 베이지안 방법론 등이 다양하게 활용되고 있다. 또한 패턴 인식으로는 분류와 클러스터링을 기본으로 지표면이나 해수 표면 분석 등에 활발히 활용되고 있다.

데이터를 기반으로 한 기계 학습은 대부분의 경우 수치 예측 모델에 비해 수행 속도가 훨씬 빠르다. 따라서 기계 학습이 좋은 성과를 내고 있는 단기 예측에서 기존 수치 모델에 비해 큰 경쟁력이 있으며, 수치 예측과는 상관없는 패턴 인식 분야에서는 독자적인 성장 가능성이 높다. 앞으로 기계 학습 기반의 기상 예측과 패턴 인식을 더욱 정교화 하고 실시간 데이터 처리를 강화함으로써 진일보된 자연재해 분석과 관리가 가능하리라 기대한다.

References

Alexandridis, A., Chondrodima, E., Efthimiou, E., Papadakis, G., Vallianatos, F., and Triantis, D. (2014). Large earthquake occurrence estimation based on radial basis function neural networks. *Geoscience and Remote Sensing, IEEE Transactions on*, **52**, 5443-5453.

Angayarkkani, K. and Radhakrishnan, N. (2010). An intelligent system for effective forest fire detection using spatial data. *International Journal of Computer Science and Information Security*, **7**, 202-208.

Arrue, B. C., Ollero, A. and Matinez de Dios, J. R. (2000). An intelligent system for false alarm reduction in infrared forest-fire detection. *Intelligent Systems and their Applications, IEEE*, **15**, 64-73.

Chakraborty, P., Marwah, M., Arlitt, M. and Ramakrishnan, N. (2012). Fine-grained photovoltaic output prediction using a bayesian ensemble. *Twenty-Sixth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, Toronto, Canada.

Chang, F. J. and Wang, K. W. (2013). A systematical water allocation scheme for drought mitigation. *Journal of Hydrology*, **507**, 124-133.

Choi, H., Park, H. W. and Park, C. (2013). Support vector machines for big data analysis. *Journal of the Korean Data & Information Science Society*, **24**, 989-998.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, **20**, 273-297.

Cortez, P. and Morais, A. (2007). A data mining approach to predict forest fires using meteorological data. *Proceedings of the 13th EPIA2007- Portuguese conference on artificial intelligence*, Guimarães, Portugal, 512-523.

Donges, J. F., Zou, Y., Marwan, N. and Kurths, J. (2009). Complex networks in climate dynamics. *The European Physical Journal Special Topics*, **174**, 157-179.

Dos Santos, J. A., Gosselin, P. H., Philipp-Foliguet, S., Torres, R. S. and Falao, A. X. (2012). Multiscale classification of remote sensing images. *Geoscience and Remote Sensing*, **50**, 3764-3775.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. and Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, **9**, 155-161.

Foley, A. M., Leahy, P. G., Marvuglia, A. and McKeogh, E. J. (2012). Current methods and advances in forecasting of wind power generation. *Renewable Energy*, **37**, 1-8.

He, H., Cao, Y., Cao, Y. and Wen, J. (2011). Ensemble learning for wind profile prediction with missing values. *Neural Computing and Applications*, **22**, 287-294.

Jursa, R. and Rohrig, K. (2008). Short-term wind power forecasting using evolutionary algorithms for the automated specification of artificial intelligence models. *International Journal of Forecasting*, **24**, 694-709.

Kalra, A., Miller, W. P., Lamb, K. W., Ahmad, S. and Piechota, T. (2013). Using large-scale climatic patterns for improving long lead time streamflow forecasts for Gunnison and San Juan River Basins. *Hydrological Process*, **27**, 1543-1559.

Kim, Y. and Cho, K. H. (2013). Big data and statistics. *Journal of the Korean Data & Information Science Society*, **24**, 959-974.

Kulahci, F., Inceoz, M., Dogrua, M., Aksoyb, E. and Baykara, O. (2009). Artificial neural network model for earthquake prediction with radon monitoring. *Applied Radiation and Isotopes*, **67**, 212-219.

Kusiak, A., Wei, X., Verma, A.P. and Roz, E. (2012). Modeling and prediction of rainfall using radar reflectivity data: A data-mining approach. *IEEE Transactions on Geoscience and Remote Sensing*, **51**, 2337-2342.

Li, X., Plale, B., Vijayakumar, N., Ramachandran, R., Graves, S. and Conover, H. (2008). Real-time storm detection and weather forecast activation through data mining and events processing. *Earth Science Informatics*, **1**, 49-57.

Magoulas, R. and Lorica, B. (2009). *Introduction to big data.* Release 2.0, 11.

Mellit, A., Massi Pavan, A. and Benghanem, M. (2012). Least squares support vector machine for short-term prediction of meteorological time series. *Theoretical and Applied Climatology*, **111**, 297-307.

Mohandes, M. A., Halawani, T. O., Rehman, S. and Hussain, A. A. (2004). Support vector machines for wind speed prediction. *Renewable Energy*, **29**, 939-947.

Moustra, M., Avraamides, M. and Christodoulou, C. (2011). Artificial neural networks for earthquake prediction using time series magnitude data or seismic electric signals. *Expert Systems with Applications*, **38**, 15032-15039.

Muttil, N. and Chau, K. W. (2007). Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Applications of Artificial Intelligence*, **20**, 735-744.

Ohashi, O. and Torgo, L. (2012). Wind speed forecasting using spatio-temporal indicators. *20th European Conference on Artificial Intelligence*, 242, 975-980.

Panakkat, A. and Adeli, H. (2008). Recent efforts in earthquake prediction (1990-2007). *Natural HazardsR eview*, **9**, 70-80.

Peters, J., Verhoest, N. E. C., Samson, R., Meirvenne, M. V., Cockx, L. and Baets, B. D. (2009). Uncertainty propagation in vegetation distribution models based on ensemble classifiers. *Ecological Modelling*, **220**, 791-804.

Petropoulos, G. P., Arvanitis, K. and Sigrimis, N. (2012). Hyperion hyperspectral imagery analysis combined with machine learning classifiers for land use/cover mapping. *Expert Systems with Applications*, **39**, 3800-3809.

Race, C., Steinbach, M, Ganguly, A. R., Semazzi, F. and Kumar, V. (2010). A knowledge discovery strategy for relating sea surface temperatures to frequencies of tropical storms and generating predictions of hurricanes under 21st-century global warming scenarios. *Conference on Intelligent Data Understanding*, Mountain View, California, USA. 204-212.

Radhika, Y. and Shashi, M. (2009). Atmospheric temperature prediction using support vector machines. *International Journal of Computer Theory and Engineering*, **1**, 1793-8201.

Rasouli, K., Hsieha, W. W. and Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, **414**, 284-293.

Sakr, G. E., Elhajj, I. H., Mitri, G. and Wejinya, U. C. (2010). Artificial intelligence for forest fire prediction.

Advanced Intelligent Mechatronics (AIM), 2010 IEEE/ASME International Conference on, 1311-1316, IEEE.

Sharma, N., Sharma, P., Irwin, D. and Shenoy, P. (2011). Predicting solar generation from weather forecasts using machine learning. *IEEE*, 528-533.

Snell, S. E., Gopal, S. and Kaufmann, R. K. (2000). Spatial interpolation of surface air temperatures using artificial neural networks: evaluating their use for downscaling GCMs. *Journal of Climate*, **13**, 886-895.

Steinbach, M., Tan, P. N., Kumar, V., Klooster, S. and Potter, C. (2003). Discovery of climate indices using clustering. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, DC, U.S.A. 446-455.

Steinhaeuser, K., Chawla, N. V. and Ganguly, A. R. (2010). Complex networks as a unified framework for descriptive analysis and predictive modeling in climate. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **4**, 497-511.

Steinhaeuser, K., Chawla, N. V. and Ganguly, A. R. (2011). Comparing predictive power in climate data: clustering matters. *Advances in Spatial and Temporal Databases, Lecture Notes in Computer Science*, **6849**, 39-55.

Storch, H. V. and Zwiers, F. W. (1999). *Statistical analysis in climate research*, Cambridge University Press, United Kingdom.

Tasadduq, I., Rehman, S. and Bubshait, K. (2002). Application of neural networks for the prediction of hourly mean surface temperatures in Saudi Arabia. *Renew Energy*, **25**, 545-554.

Tsonis, A. A., Swanson, K. L. and Roebber, P. J. (2006). What do networks have to do with climate?. *BAMS*, **87**, 585-595.

Vatsavai, R. R., Bright, E., Varun, C., Budhendra, B., Cheriyadat, A. and Grasser, J. (2011). Machine learning approaches for high-resolution urban land cover classification: A comparative study. *2nd International Conference on Computing for Geospatial Research & Applications*, **11**, 1-10.

Vega-Garcia, C., Lee, B. S., Woodard, P. M. and Titus, S. J. (1996). Applying neural network technology to human-caused wildfire occurrence prediction. *AI Applications*, **10**, 9-18.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G. A. and Torricelli, P. (2011). Application of a random forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. *Ecological Modelling*, **222**, 1471-1478.

Voyant, C., Muselli, M., Paoli, C. and Nivet, M. L. (2011). Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. *Energy*, **36**, 348-359.

Journal of the Korean Data & Information Science Society
2015, **26**(5), 1105–1115

http://dx.doi.org/10.7465/jkdi.2015.26.5.1105
한국데이터정보과학회지

Big data mining for natural disaster analysis[†]

Young-Min Kim[1] · Mi-Nyeong Hwang[2] · Taehong Kim[3] ·
Chang-Hoo Jeong[4] · Do-Heon Jeong[5]

[12345]Disaster Information Service Lab., Korea Institute of Science and Technology Information

Received 5 August 2015, revised 7 September 2015, accepted 16 September 2015

Abstract

Big data analysis for disaster have been recently started especially to text data such as social media. Social data usually supports for the final two stages of disaster management, which consists of four stages: prevention, preparation, response and recovery. Otherwise, big data analysis for meteorologic data can contribute to the prevention and preparation. This motivated us to review big data technologies dealing with non-text data rather than text in natural disaster area. To this end, we first explain the main keywords, big data, data mining and machine learning in sec. 2. Then we introduce the state-of-the-art machine learning techniques in meteorology-related field sec. 3. We show how the traditional machine learning techniques have been adapted for climatic data by taking into account the domain specificity. The application of these techniques in natural disaster response are then introduced (sec. 4), and we finally conclude with several future research directions.

Keywords: Big data, data mining, machine learning, meteorologic data, natural disaster.

[†] This work was supported by the IT R&D program of MSIP/IITP (B010-15-0353, High performance database solution development for integrated big data monitoring and analytics).

[1] Senior researcher, Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea.

[2] Senior researcher,Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea.

[3] Senior researcher,Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea.

[4] Senior researcher, Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea.

[5] Corresponding author: Senior researcher, Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea. E-mail: heon@kisti.re.kr