클러스터링 기법을 이용한 산불 데이터의 상관관계 분석

Correlation Analysis of forest fire data based on Clustering Method

김은희, 지정희, 손호선, 류근호, 이충호*
충북대학교 데이터베이스/바이오인포메틱스 연구실
{ehkim, jhchi, hsshon0621, khryu}@dblab.cbu.ac.kr
한국전자통신연구원 텔레메틱스/USN 연구단 공간정보연구팀
leech@etri.re.kr*

Eun Hee Kim, Jeong Hee Chi, Ho Sun Shon, Keun Ho Ryu, Chung Ho Lee*
Database/Bioinformatics Laboratory of Chungbuk National University
Spatial Information Research Team Telematics/USN Research Division, ETRI*

요 약 이 논문에서는 산불 발생의 패턴을 예측하기 위해 데이터 마이닝의 클러스터링 기법을 이용하여 산불 데이터를 그룹화하고 그 결과를 이용하여 산불 데이터의 상관관계를 분석하는 방법을 제안하였다. 즉, 클러스터링 기법을 이용하여 산불 데이터를 사용자가 원하는 수의 그룹으로 분류하고, 생성된 산불 데이터 클러스터 모델을 이용하여 새로운 유형의 산불 패턴을 예측 할 수 있도록 하였다. 또한 결과 클러스터의 생성을 위해 이전의 산불 분포 데이터를 저장 관리하여 클러스터 간의 상관관계 분석을 통해 시퀀스를 생성하였고, 생성된 각각의 클러스터 시퀀스를 통합하여 클러스터들의 시퀀스를 추출하여 산불이 발생한 이후의 향후 발생 가능한 산불 유형을 예측하기 위한 방법을 제공하였다. 이는 과거에 발생된 산불의 유형뿐만 아니라 새로운 형태의 산불 유형 분류나 분석에 이용 가능하다.

1. 서 론

최근 지구 온난화로 인한 기상이변 등으로 건조한 날씨가 계속되면서 산불이 많이 발생하고 있다. 산불은 과거에서 오늘에 이르기까지 인위적 산림피해 가운데 가장 큰 물리적인 피해를 일으켰다. 또한 산불은 그 강도와 지속기간에 따라서 정도의 차이는 있으나 토양의 물리, 화학적 성질을 변화시키고 호우시 토양침식이 용이하도록 토양에 작용한다. 따라서 정확한 산불 위험 예측을 위해서 산불발생 원인분석을 통하여 산불 위험을 예측할 수 있는 연구들이 진행되었다. 이들 연구는 대형 산불 방지 및 산불로 인한 피해를 최소화하기 위해 지역별 온도, 습도, 풍속 등의 기상조건과 지형 및 임상조건을 종합 분석하여 실시간 및 24 ~ 48시간까지 산불경보를 발령함으로써 산불 발생 가능성이 높은 지역에 인력을 집중 배치하고 해당 지역 산불 담당자 및 주민에게 알려 사전에 대처할 수 있도록 하는 시스템이다. 또한 이들 시스템들은 산불발생에 영향을 미치는 요인인 임상과 지형과 같은 지역적인 정보와 온도, 습도 등과 같은 기상 정보를 모두 분석하여 산불 위험률을 예측하고 있다. 하지만, 이들 시스템들은 산불 발생 위험률을 예측하여 지역 주민이나 관공서에 경보를 발생시킴으로서 대비하는 데 목적이 있다. 이 논문에서는

이전에 발생된 산불 데이터를 분석하여 산불 발생 패턴과 산불의 진행 경로를 예측하기 위해 데이터 마이닝 기법 중 클러스터링 기법을 기반으로 산불 데이터의 상관관계를 분석한다. 클러스터링 기법을 이용하여 산불 데이터를 사용자가 원하는 수의 그룹으로 분류하고, 생성된 산불 데이터 클러스터 모델을 이용하여 새로운 유형의 산불 패턴을 예측 할 수 있도록 한다. 또한 결과 클러스터의 생성을 위해 이전의 산불 분포 데이터를 저장 관리하여 클러스터 간의 시퀀스를 생성하며, 생성된 각각의 클러스터 시퀀스를 통합하여 클러스터들의 시퀀스를 추출하여 산불이 발생한 이후의 향후 발생 가능한 산불 유형을 예측하기 위한 방법을 제공한다.

이 논문의 구성은 다음과 같다. 2장에서는 현재 서비스 되고 있는 산불 위험률 예측 시스템의 분석과 문제점을 기술하고, 3장에서는 이 논문에서 제안한 클러스터링 기법을 기반으로 한 산불 데이터의 상관관계 분석 방법에 대해서 기술한다. 4장에서는 제안한 방법의 실험 평가를 통해 유용성을 검증한다. 마지막으로 5장에서 결론 및 향후 연구를 끝으로 맺는다.

2. 관련연구

우리나라에서는 최근 5년간 산불 발생의 원인은 캐나다나 미국과 같이 자연적 원인에 의한 산불 발생은 거의 없고 대부분이 등산객 및 논. 밭두렁 소각 부주의로 발생하는 인간 활동과 관련되어 있어 사전에 산불 위험지역을 예측하여 조기에 차단하기 위한 산불 위험율 예측 시스템을 구축하여 서비스를 제공하고 있다[3]. 이 시스템은 과거 5년 동안의 기상자료를 기상관측소로부터 획득하고, 이 기간 동안의 산불 발생 현황자료를 이용하여 산불발생 유무와 기상변수를 로지스틱회귀모형[6]에 적용하여 산불발생확률모형을 추정하였다. 미국의 산불위험율 시스템(NFDRS: National Forest fire Danger Rating System)[2] 개발은 1940년과1945년에 유타주에서 미국 농무성 산림청에 의해 소집된 산불방지위원회에서 국가 전체에 일률적인 산불 위험등급시스템 구축이 필요하다고 강조된 후부터 시작되었다. 이 시스템은 화재제어 활동 계획에 도움을 주는 4개의 지수를 제공한다. 이들 지수는 인위적 산불발생지수, 번개에 의한 화재발생지수, 연소지수 그리고 연소물량지수로 구분된다. 또한 산불위험율등급, 화재발생등급, 발화가능성 등급 그리고 연소물량지수는 0에서 100등급으로 위험구간을 표시하고 있다. 캐나다에서는 1925년부터 산불발생위험율에 관한 연구를 시작하여 1980년 캐나다 산불 위험율 시스템(CFFDRS: Canadian Forest Fire Danger Rating System)[1]을 개발하였다. 이 시스템은 실시간 산불 발생 모니터링, Mapping, Spatial Fire Management System을 구축하고 산불확산 및 진화전략에 활용하고 있다. 일본에서는 건조하고 바람이 강하게 부는 상황하에서는 화재가 급격하게 확대되고 넓은 범위가 소실되는데, 이를 제압하기 위해 필요한 많은 시간과 인원 및 경비를 줄이기 위해 과거의 데이터나 온라인 기상정보에 근거하여 시시각각으로 발생하는 산불화재의 발생위험도 지수를 산출함으로서 발생과 확대의 위험도를 예측하고 화재가 늘어나는 것을 방지하기 위해 산불발생위험도 확산예측 시스템을 연구하고 있다. 이 시스템은 지형과 산림에 관한 데이터베이스나 기상정보와 종합하여 시시각각 발생하는 산불화재의 발생위험도 지수를 산출하여 일본의 산불위험지도와 산불 확산도를 예측하는 서비스를 제공하고 있다.

하지만 이들 연구는 기상관측소나 현장 실시를 통하여 산불 위험률 예측을 위해 필요한 데이터를 수집하고 있다. 그 결과로서, 과거의 기상정보와 지형정보를 분석하여 산불 위험도 예측 즉, 산불의 발생 여부를 분석하는데 그 목적이 있다. 산불은 발생 여부를 사전에 예측하는 것도 중요하지만, 산불이 발생 한 후 산불의 진행 경로를 예측한다면 많은 피해를 줄 일 수 있다. 따라서 우리는 이 논문에서 데이터 마이닝 기법 중 클러스터링 기법을 이용하여 산불 데이터의 유사성을 분석하여 분석된 결과를 토대로 산불의

발생의 패턴 및 진행 경로를 예측하는 방법을 제공한다.

클러스터링 기법[4]은 잠재적인 데이터에서 그룹들을 탐사하거나 관심 있는 분포를 확인하는 데 유용한 방법이다.

3. 산불 데이터의 상관관계 분석

이 장에서는 산불 데이터를 데이터베이스에 저장하기 위한 스키마를 설계한다. 또한 산불 데이터에 클러스터링 분석을 적용하기 위해 산불 데이터간의 유사성을 정의하는 방법에 대해 기술한다. 그리고 저장된 산불 데이터와 정의된 유사성을 기반으로 산불 데이터간의 유사성을 분석한 후, 그 결과를 이용하여 산불의 진행 경로를 예측하는 패턴을 추출하는 과정에 대해 기술한다.

3.1 산불 데이터베이스 스키마

산불 데이터를 구조적으로 저장 관리하기 위하여 이 논문에서는 관계형 데이터베이스를 이용한다. 산불 클래스 스키마는 발생한 산불을 식별할 수 있는 FID(fire identifier)와 해당 산불의 영향(impact)으로 구성된다.

Impact는 0~5의 숫자 값으로 표현되며, 각 숫자 값의 의미는 <표 1>과 같다.

<표 1> Impact 값의 정의

값	설명
0	전혀 영향 없음 (0%)
1	산불 발생 위험도 낮음(20% 미만)
2	산불 발생 위험도 미약(20~40%)
3	산불 발생 위험도 중간(40~50%)
4	산불 발생 위험 큼(50~70%)
5	산불 발생 위험 매우 큼(70%이상)

산불 클래스의 하위 클래스로서 발화지 클래스, 점화지 클래스, 시간 클래스로 구성되어 있다. 발화지와 점화지 클래스는 산불이 초기에 발화된 지역과 마지막으로 진화된 지역을 의미하는 것으로, 풍향, 습도등과 같은 산불 발생 당시의 주변 정보를 포함하고 있다. 시간 클래스는 산불 발화 시간과 발화되어 진화된 시간의 정보로 구성되어 있다. 표 2는 클러스터링에 사용될 산불 데이터의 스키마를 보여준다. FID는 각 레코드를 구분하기 위한 필드로서 클러스터링에는 사용되지 않는다.

<표 2> 산불 데이터 스키마

속성명	설명
FID	산불 데이터의 식별자
Impact	산불 발생 위험도 평가
Source	산불 발생지
Destination	산불 진화지
Dir_wind	바람의 방향
speed	바람의 속도
precipitation	강수량
hummidity	습도
temperature	온도
timestamp	시간정보

3.2 클러스터링을 이용한 데이터 추상화

산불 데이터간의 상관간계를 분석하고, 이를 이용하여 유사한 산불 데이터를 그룹화하기 위하여 이 논문에서는 데이터 마이닝 기법 중 클러스터링을 이용한다. 산불 데이터간의 유사성은 산불 데이터 간의 근접지수에 의한 방법을 이용하여 클러스터링 한다. 근접 지수는 두 데이터 개체간의 유사성이나 연관성을 측정할 수 있는 함수로서 주로 거리 개념을 이용하여 측정한다. 속성간의 유사성을 정의하기 위해서 유클리드 거리 함수를 이용한다. 이는 동일한 속성 값들을 가지는 데이터 개체는 유사하다는 가정을 기반으로 한다. n개의 속성을 가지는 데이터 개체 x, y가 주어지고 각 속성의 값이 x_i, y_i $(0 \le i \le n)$로 정의될 때 두 개체간의 거리를 추출하기 위한 함수는 아래 식과 같이 정의된다.

$$dis(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

저장된 산불 데이터와 유사도 측정 함수를 이용하여 클러스터링을 수행하기 위한 절차는 두 단계로 이루어진다. 첫 번째 단계에서는 데이터 전처리과정이고, 두 번째 단계에서는 클러스터링 과정이다.

● 단계 1: 데이터 전처리

입력 데이터 집합에 대해 클러스터링 기법을 수행 할 수 있도록 적당한 데이터 처리를 하는 과정으로서, 이 과정에서는 크게 2 가지 작업을 수행한다. 첫 번째는 클러스터링의 효율적인 수행을 위하여 적당한 속성들을 선택하고 필요에 따라 확장된 속성을 추가하는 과정이다. 이 논문에서는 클러스터링을 효율적으로 수해하기 위해 연관규칙 기법을 이용하여 속성들을 선택 한다. 연관규칙 기법은 속성들 간의 연관성을 분석하는 데 많이 사용되는 데이터 마이닝 기법 중에 하나이다.

단계 1에서 수행되는 두 번째 작업은 입력 데이터의 정규화이다. 산불 데이터의 특성상 속성의 값이 유사한 경우가 많기 때문에 각각의 데이터에 대해서 모두 클러스터링을 수행하는 것 보다는 데이터의 분포를 고려하여 클러스터링을 수행하는 것이 클러스터링 비용이나 유지 면에서 더 효율적이다. 이 논문에서는 원래의 데이터 개체를 직접 클러스터링에 사용하지 않고, 데이터 개체의 분포를 이용하여 클러스터링을 수행한다. 즉, 어떤 속성 i 와 그 속성 값의 m 개 레코드가 주어졌을 때, 속성 i의 평균값 *avg[i]* 과 표준 편차 *std[i]*를 계산하여 클러스터링에 이용한다.

평균과 표준 편차를 계산하는 식은 다음과 같이 정의된다.

$$avg[i] = \frac{1}{m}\sum_{j=0}^{m} inst[i]_j$$

$$std[i] = \frac{1}{m-1}\sum_{j=0}^{m}\sqrt{(inst[i]_j - avg[i])^2}$$

계산된 평균과 표준편차를 이용하여 각 데이터 개체의 속성 값들, *inst[i]*, 은 다시 다음 식에 의해 변경된다.

$$inst[i] = \frac{inst[i] - avg[i]}{std[i]}$$

● 단계 2: 클러스터링 수행

단계 1에서 처리된 입력 데이터 집합에 대해 실제 클러스터링을 수행하는 과정이다. 주어진 데이터 집합으로부터 클러스터를 생성하기 위하여 CURE 알고리즘[5]을 기반으로 수행하였다. 이 알고리즘은 고 차원의 데이터 클러스터링을 지원하는 알고리즘으로서 산불 데이터의 클러스터링에 적당한다. 이 알고리즘에서는 최초에 데이터 개체 각각을 분리된 클러스터로 간주하고, 모든 개체들이 사용자가 입력한 k 개의 클러스터를 형성할 때까지 거리함수에 의해 클러스터들을 통합함으로서 클러스터링을 수행하는 상향식의 계층적인 클러스터링 기법이다. CURE는 클러스터의 합병을 위해서 거리를 계산할 때 c 개의 대푯값을 이용하는데, 이 값들은 각 클러스터로부터 클러스터의 모양을 가장 잘 표현 할 수 있는 c 개의 적절히 분포된 포인트를 선택하여 클러스터의 중심점으로 사용자가 입력한 수축률만큼 축소시켜 구해진 점들이다. CURE에서 두 클러스터 *u, v* 간의 거리, *dist-cluster(u,v)*는 다음 식과 같이 정의된다. 이때 *p*와 *q* 는 각 클러스터의 대표값들이다.

$$dis-cluster(u,v) = \min_{p\in u,\, q\in v} dis(p,q)$$

위의 식에 의해 정의된 클러스터간의 거리는 병합하는 과정에서 클러스터간의 유사도를 판단하는 기준이 된다. 이 단계의 결과로서 우리는 유사한 데이터 개체끼리 그룹 화된 몇 개의 데이터 집합, 클러스터를 얻는다.

3.3 산불 패턴 추출

이 단계는 생성된 클러스터에 대해 그 생성 원인을 분석하여 클러스터간의 순차적인 연관관계를 추출하는 과정이다. 클러스터간의 연관관계를 추출하기 위해 우리는 생성된 클러스터에 포함된 이전 데이터의 분포를 이용한다. 특정 클러스터에 포함된 산불 데이터의 이전 분포를 분석함으로서 해당 클러스터에 포함된 산불 유형 이전에 자주 발생하는 산불 유형을 탐사할 수 있다. 즉, 산불이 크게 번지기까지의 과정을 유추해 낼 수 있다. 두 클러스터 a, b 가 주어졌을 때, 위와 같은 분석 과정을 통해 b의 이전 클러스터 *(cluster_pre)*가 a 이면, a의 다음 클러스터

*(cluster_post)*는 b로 정의되고, 두 클러스터의 관계는 a->b 로 표현된다. 이러한 각 클러스터간의 순차적인 관계의 분석을 통해 얻을 수 있는 최종 결과는 클러스터로 이루어진 클러스터 패턴이다.

4. 실험 및 평가

이 장에서는 클러스터링의 성능에 대한 실험과 클러스터의 시퀀스를 추출하기 위한 실험을 수행하고 그 결과에 대해서 설명한다.

4.1 실험 환경 및 데이터

실험 환경은 Windows기반 PentiumIV, 512MHz 기반에 JAVA언어로 구현하였고, 데이터베이스는 MySQL을 사용하였다.

실험을 위해 사용한 실험 데이터는 한 지역의 1년간의 기후 데이터를 사용하였다. 이 데이터 집합은 약 365개의 데이터 인스턴스로 구성되어 있다.

이 논문에서 제안한 방법에 대해 두 가지 측면에 대해 평가를 수행한다. 첫 번째는 클러스터링의 성능을 평가하기 위한 실험이다. 이 실험은 생성된 각 클러스터의 정확도를 평가하는 것이다. 두 번째는 생성된 클러스터간의 상관관계를 분석하여 이를 기반으로 클러스터의 패턴을 생성할 수 있는지의 여부를 평가하기 위한 실험이다.

4.2 클러스터링 성능

입력 데이터에 대해 클러스터링을 수행하기 위해서는 먼저 사용자 입력 변수를 결정하여야 한다. 이 논문에서는 사용자가 입력하는 변수로서 선택된 포인터를 대표 값으로 변환하기 위한 수축률과 클러스터의 대표 값의 개수이다. 실험을 위한 수축률과 대표 값의 개수를 선택하기 위하여 우리는 임의로 선택된 10%의 트레이닝 데이터 집합에 대해 클러스터링을 수행하였다.

결정된 입력 변수와 전체 트레이닝 데이터 집합을 이용하여 클러스터를 생성한 후, 트레이닝 데이터 집합에 포함되지 않은 새로운 데이터 집합을 이용하여 클러스터의 결과를 분석하였다. 트레이닝 데이터 집합을 이용하여 형성한 모델에 대해 테스트 데이터 집합을 이용하여 실험한 결과 <표 3>과 같이 생성되었다.

<표 3>클러스터링 결과

CID	속성 분포도
1	42%
2	25%
3	33%

<표 4> 클러스터내 중심 값

속성 \ CID	1	2	3
기온	9.38	26.87	20.23
습도	54.88	77.40	63.63
강수량	63.44	230.57	98.25
풍속	337	189	250
풍향	5	5	6
영향도	1	2	3

4.3 클러스터링 패턴 분석

이 실험은 생성된 클러스터에 대해 각 클러스터의 이전 클러스터를 정의하고 이를 기반으로 클러스터의 패턴을 생성할 수 있는지의 여부를 평가하기 위한 실험이다. <표 4>는 속성 값의 각 클러스터의 중심 값을 나타낸다.

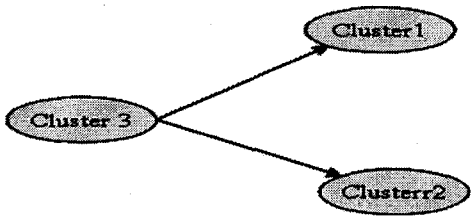
클러스터링 결과를 바탕으로 하여 클러스터간의 상관관계를 분석하였다. 분석된 결과는 <표 5>와 같다.

<표 5> 클러스터간의 상관도 분석

	클러스터1	클러스터2	클러스터3
클러스터1	-	225.057	94.763
클러스터2	225.057	-	146.510
클러스터3	94.763	146.510	-

<표 5>의 상관계수 결과를 토대로 클러스터1은 클러스터 3과 가장 근접해 있고, 클러스터 2도 역시 클러스터 3과 근접해 있고,

클러스터 3은 클러스터 1과 근접해 있다는 것을 알 수 있다. 이 결과를 토대로 <그림 1>과 같은 클러스터간의 패턴을 생성해 낼 수 있다.



<그림 1> 클러스터 시퀀스 패턴

추출된 시퀀스 패턴을 통해서 산불의 유형이 클러스터 3에 속한다면 다음의 산불 유형은 클러스터 1로 번질 확률이 더 높다는 것을 예측 할 수 있다. 그것은 상관도 분석 결과 클러스터 3에서는 중심 값이 클러스터 1이 2보다 더 가깝기 때문이다.

5. 결 론

이 논문에서는 산불 발생의 패턴을 예측하기 위해 데이터 마이닝의 클러스터링 기법을 이용하여 산불 데이터를 그룹화하고 그 결과를 이용하여 산불 데이터의 상관관계를 분석하는 방법을 제안하였다. 산불 데이터의 상관관계 분석을 위해 우리는 두 단계로 작업을 수행하였다. 첫 번째 단계에서는 데이터 전처리과정이고, 두 번째 단계에서는 클러스터링 과정이다. 데이터 전처리 과정에서는 입력된 데이터 중에서 클러스터링을 위해 적절한 속성을 선택하고, 선택된 속성들 중에서 클러스터링에 적합하게 변환하는 작업이 수행된다. 전처리과정이 완료되면, 클러스터링 작업을 수행하게 되는데 이 논문에서는 CURE 알고리즘을 기반으로 산불 데이터를 클러스터링 하였다. 우리는 결과 클러스터의 생성을 위해 이전의 산불 분포 데이터를 저장 관리하여 클러스터 간의 시퀀스를 생성하였고, 생성된 각각의 클러스터 시퀀스를 통합하여 클러스터들의 전체 시퀀스를 추출하므로서 산불이 발생한 이후의 향후 발생 가능한 산불 패턴 유형을 예측하기 위한 방법을 제공하였다. 이 결과를 토대로 과거에 발생된 산불의 유형뿐만 아니라 새로운 형태의 산불 유형 분류나 분석에도 이용 가능하다. 또한 생성된 클러스터간의 생성 원인의 분석에 의한 클러스터 간의 순차적인 관계의 추출을 통해 산불 경로를 예측하여 산불 진압을 위한 전략을 세우는데 도움을 줄 수 있다.

Acknowledgement

이 연구는 한국전자통신연구원 텔레메틱스/USN 연구단 공간정보연구팀의 연구비 지원에 의해 수행되었음.

< 참 고 문 헌 >

[1] CFFDRS, http://fire.cfs.nrcan.gc.ca

[2] USFS, http://www.fireplan.gov

[3] 산림청, http://www.foa.go.kr

[4] Periklis Andritsos, data clustering techniques, qualifying oral examination paper, 2001

[5] Sudipto Guha, Rajeev Rastogi and Jyuseok Shim, CURE: An Efficient Clustering Algorithm for Large Database, In proceedings of the international conference on management of data, SIGMOD Record, Seattle, WA, USA, 14, ACM Press, Vol.27(2), pp.73-84, Jun, 1998.

[6] Hosmer DW, Lemeshow S: Applied Logistic Regression. New York: John Wiley &Sons,1998.