기상자료 군집화를 통한 지형적 특성 연구

김민진, 이일병
연세대학교 컴퓨터과학과 인공지능연구실
mjkim@csai.yonsei.ac.kr, yblee@csai.yonsei.ac.kr

Clustering Weather Data for Study of Local Distinction

Minjin Kim
University of Yonsei, Computer Science, AI Lab

요 약

매일 쏟아져 나오는 방대한 양의 기상자료는 현재의 대기상태를 대표하기도 하지만 그 지역의 지형적 특성을 나타내고 있다. 이번 연구는 수원지역의 일일 기상자료를 토대로 지형적 특성과 그에 따른 기상현상(바람, 안개)알고자 한다. K-means를 이용 특정 기상현상끼리 군집화하여 지형적 특성과 비교하였다.

1. 서 론

매 순간 쏟아져 나오는 기상데이터는 현재의 대기특성을 대표하기도 하지만 지형의 특이점(하천, 산)에 영향을 받기도 한다. 현재의 수치모델은 중규모로 방대한 양의 데이터와 광대한 지역을 계산하기 때문에 지형적 작은 규모의 예보 및 지형적 특성이 반영된 데이터는 무시되어지기도 한다.

데이터 마이닝(data mining)은 대량의 데이터로부터 패턴 인식, 통계적 기법, 인공지능 기법 등을 이용하여 데이터간의 상호 관련성, 패턴, 추세 등 의사결정에 유용한 정보를 추출해 내는 과정으로서 지식탐사의 핵심적인 역할을 담당한다(Bruce, 1996). 데이터 마이닝은 문제의 영역이나 그 목적에 따라서 다양한 방법들이 존재하는데, Cooley(1997)는 군집분석(clustering analysis), 분류규칙 발견(classification rule discovery), 연관규칙 발견(association rule discovery), 연속패턴 발견(sequence pattern discovery), 시각화(visualization) 등의 다섯 가지로 분류하였다.

이번 연구는 K-means을 이용하여 군집화하였고 이를 통해 지형적 특성과 결합, 특정 지역의 기상현상에 미치는 영향이 중규모적 특성이 강한지, 그지역의 지형적 특성이 영향이 큰지를 분석하였다. 연구는 수원 지역의 1994년~2005년의 가을(9월, 10월)의 지상관측자료를 이용하였으며, SPSS Clementine 10을 이용하여 클러스터링하였고 분석하였다.

2. 군집화

군집화/세분화(Clustering/Segmentation)작업은 전체를 어떤 측정기준에 따라 유사한 그룹으로 나누는 것이다. 하나의 굽집은 유사성에 따라 모인 일련의 대상 객체이다. 유사성에 따른 군집기법은 계량적인 방법으로 유사성을 측정하는 강력한 기법이다. 스스로 발견하고 이를 학습해 가는 무감독학습(Unsupervised Learning) 유형으로 훈련 집합내에 있는 관련된 대상으로 군집을 발견하고 나서 각 군집에 대한 설명을 찾게 된다. 분류작업과의 차이점은 군집은 클래스가 미리 정해져 있지 않고, 전체를 유사성에 따라서 분류하여 군집화 시키는 것이다.

K-means는 군집화 문제를 해결하는 가장 간단한 자율학습(Unsupervised Learning) 알고리즘 중 하나이다. 사전에 정해진 어떤수의 클러스터 수를 통해서 주어진 데이터의 집합을 분류하는 간단하고 쉬운 방법이다. K-means는 partitonal clustering에 속한다.

data 이외에 cluster의 수 k를 input으로 하며 이때 k를 seed point라고 한다. seed point는 임의로 선택되며 바람직한 cluster 구조에 관한 어떤 지식들이 seed point를 선택하는데 사용될 수 있다.

1, 처음에는 k cluster로서 시작한다. 남아있는 n-k sample들에 대해서는 가장 가까이 있는 centroid를 찾는다. 이것에 가장 가까이 있는 centroid를 가지는 것이 확인된 cluster에 sample을 포함시킨다. 각각의 sample들이 할당된 후에 할당된 cluster의

centroid가 다시 계산된다.

2. 그 data를 두 번 처리한다. 각 sample에 대하여 가장 가까이 있는 centroid를 찾는다. 가장 가까이 잇는 centroid를 가진 것으로 확인된 cluster에 sample을 위치시킨다.(이 step에서는 어떤 centroid도 다시 계산하지 않는다.)

K-means 알고리즘이 주된 결점은 결과가 초기 cluster centroid에 너무 sensitive 하다는 것과 local optima에 빠진다는 것이다. 또한 각 군집에 대한 분석이 모호하다는 단점을 내포하고 있어 연관규칙(CART)을 이용하여 각 군집을 설명하고자 하였다.

K-means는 기상분야에서 다양하게 이용되고 있다. 관측자료를 통해 특이기상(뇌우, 토네이도)의 특징적 현상을 군집화하기도 하며, 퍼지수를 도입하여 직접적 예보를 하기위한 알고리즘으로 사용되기도 한다.[1][2] 또한 방대한 양의 격자점 데이터 및 일기도 관련 데이터를 효율적으로 저장 및 검색 하기위해 데이터들의 유형을 찾아서로 유형이 비슷한 데이터들을 하나의 클러스터로 연관지어 놓아 효율적인 저장과 검색을 할 수 있기 위해 SOM(Self Organizing Map) 기법을 사용하기도 한다.[3]

3. 실험과정

실험에 사용된 데이터는 10년간(96년~05년)의 가을 지상관측자료를 이용하였다. 이는 지형적 특이기상인 안개발생이 가을철에 빈번하기 때문이다. 가을철 안개는 대기가 안정한 상태에서 지형적으로 습기의 유입으로 인해 주로 발생하는 복사무가 주를 이루기 때문이다.

사용되어진 데이터는 다음과 같다.[4]

Type	Attribute	Units
cloud ceiling and visibility	cloud amount(s) cloud ceiling height visibility	horizontal visibility
wind	wind direction wind speed	degrees from true north knots
precipitation	precipitation type precipitation intensity	nil, rain, etc. nil, light, moderate, heavy
spread and temperature	dew point temperature dry bulb temperature	degrees Celsius degrees Celsius
pressure	pressure trend	kiloPascal × hour -1

대부분의 데이터는 연속형으로 이루어져 있다. dataset을 발견하고자 하는 결과와 관계가 없거나, 연관이 있다 하더라도 매우 미비한 영향을 주는 자료는 배제하였다. 이번 연구는 새벽에 발생하는 복사무를 대상으로 연구하기 때문에 현상이 발생하는 06~10시 사이의 데이터는 사용하지 않았다.

4. 실험 결과

각 클러스터에서 시간자료는 4시, 8시, 13시, 15시, 18시로 구분이 되었다. 이 시간으로 풍향과 해면기압의 분포를 조사하였을 때, 일몰 전후(17~18시)로 풍향의 분포가 바뀌는 것으로 나타나고 있다. 이는 수원의 지형적인 영향이 작용한 것으로 주간에 가열된 대지가 일몰이 일어나면서부터 해양과 대지사이의 기온역전이 일어나기 때문에 온도차로 인한 풍향의 변화라고 볼 수 있다.

그림 1 시간에 따른 풍향 분포

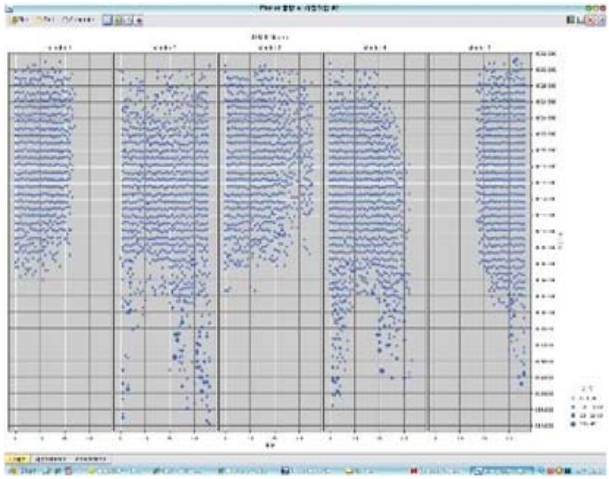
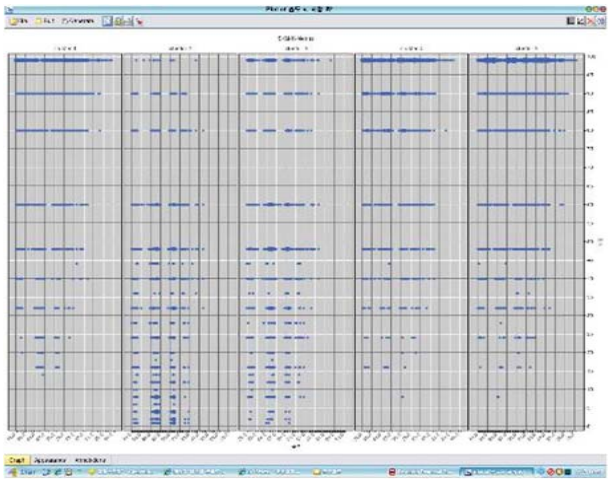


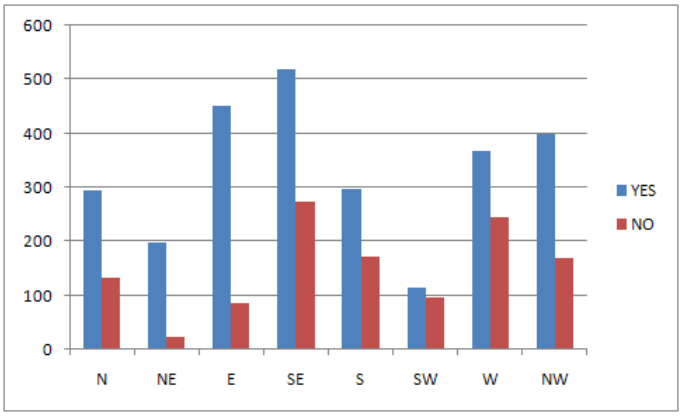
그림 2 습도에 따른 시정 분포



시정 분포의 경우 가시거리가 3200m 이하의 시정은 주로 일출 후 1~2시간 사이에 나타나고 있으며, 53%이상의 습도에서 가시거리가 줄어듬을 알 수 있다. 야간 풍속 또한 5kts내외의 미풍으로 대기가 안정화된 상태에서 대지인접한 대기층이 기온역전 현상을 띠기 때문이다. 또한 같은 습도라 하더라도 오후시간에는 안개발생 현저히 줄어듬을 알 수 있다. 주간에는 대지의 복사열로 인해 지층주변의 공기가 순환하기 때문에 안개가 소멸되게 된다.

안개가 발생하는 시간인 4시와 8시 군집에 따른 연관규칙으로 분석을 하였다. 안개발생에 대한 요인을 분석하고자 target data를 시정자료로 하고 바람 자료를 input data로 사용하였다.

도표 2. 풍향에 따른 안개 발생



안개발생의 중규모적인 요인(저기압 이동으로 인한 강수현상)을 배제하고 고기압권에서 지형적인 특성만으로 안개가 발생하는 기전을 분석하였다. 앞서 군집화 과정에서 해양의 영향으로 일몰전후의 풍향이 변화되는 것을 보았을 때, 안개의 발생은 해양의 습윤한 공기의 유입이 주된 영향으로 추정하였다. 그러나 해당군집을 분석하였을 때, 안개가 발생한 날의 주로 남동풍으로 나타났다. 이 현상을 지형적 특성으로 판단하였을 때, 관측지점의 지형적 특성이 나타나고 있음을 군집분석으로 알 수 있다. 관측지점은 남동쪽에 하천이 존재하여 남동풍이 붐으로 인해 하천의 수증기가 관측지점으로 유입되어 안개와 같은 기상현상을 발생시킨다는 것을 알 수 있다.

5. 결 론

기상데이터와 같이 짧은 시간 안에 수 기가의 자료가 생산되는 경우 이를 통해 유용한 정보를 추출하려는 시도는 매우 오래된 연구과정이다. 역학적인 모델이 들어간 수치분석이 존재하나, 이는 분석 비용 및 시간이 소모된다. 국지적인 특성을 분석하기 위해 의사결정나무나 신경망등도 많이 활용되고 있다. 이번 연구는 군집화를 통해 전체를 어떤 측정기준에 따라 유사한 그룹으로 나누고, 각 군집에 대한 분석을 실시하였다. 이를 통해 한 지역의 기상현상은 중규모 이상의 대기적인 흐름도 영향을 받게 되지만, 특정 시간의 기상(일몰 전후의 풍향변화)과 국지 기상현상(안개)은 지형적인 영향에 의해 발생함을 알 수 있게 되었다. 이로 인해 국지적인 기상분석 및 예보에 군집화같은 방법이 유용함을 알 수 있다. 향후 SOM과 같은 신경망을 이용한 군집화 및 퍼지추론을 혼합한 방법으로 기상분석[5] 및 예보생성[6]에 관한 연구

가 이루어져야 할 것 이다.

참 고 문 헌

[1] 최재훈, 이상훈 “데이터마이닝 기법.을 적용한 안개 예보칙 작성 방안 연구”, 데이터 베이스 연구, Vol 19, 제 4호

[2] DUSTIN FABBIAN AND RICHARD DE DEAR “Application of Artificial Neural Network Forecasts to Predict Fog at Canberra International Airport”, WEATHER AND FORECASTING, Vol 22, pp 372-381, 2007

[3] 이기범, 이성환, 정창성, 황치정 저 "NCEP 일기도 데이터 클러스터링을 위한 특징 벡터 추출". 한국정보과학회 2001년도 봄 학술발표논문집 제28권 제1호(B), 2001.

[4] 공군 제73기상전대.「국지예보(제6권)」, 2000.

[5] Lippmann, R. P “An introduction to computing with neural nets” IEEE Acoustics, Speech Signal Process. vol 4, pp4-22.

[6] Gardner, M. W and S. R. Dorling, "Artificial neural networks(the multilayer perceptron)--A review of applications in the atmospheric sciences". *Atmos. Enviro*n., vol 32, pp2627-2636, 1998